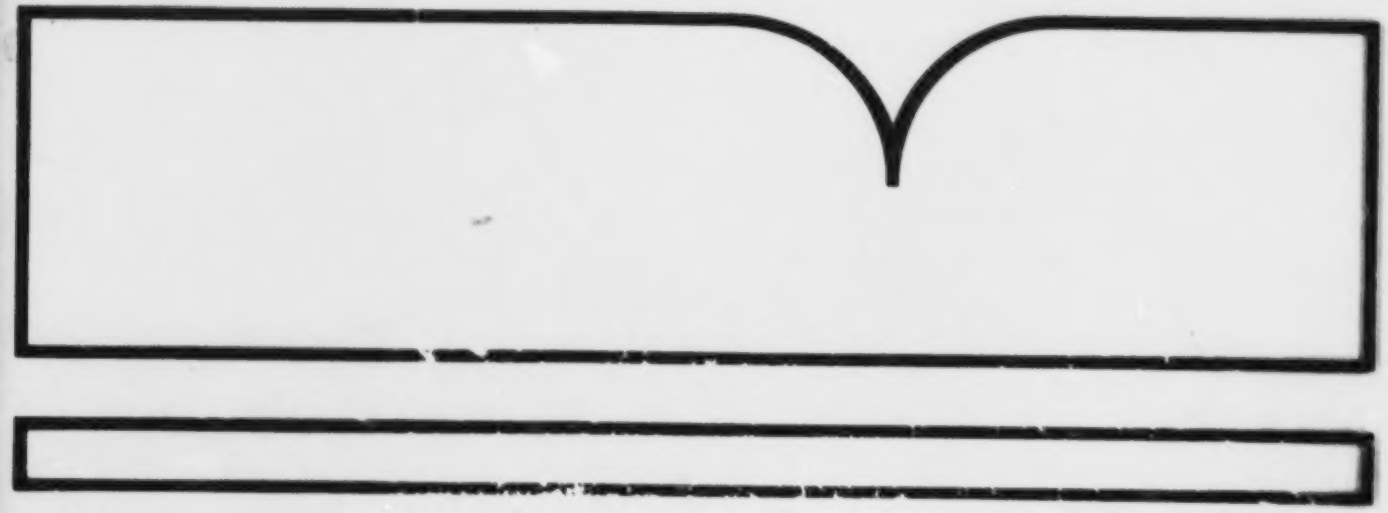
PB85-242477

Interim Procedures for Evaluating Air Quality Models: Experience with Implementation

(U.S.) Environmental Protection Agency
Research Triangle Park, NC

Jul 85



U.S. Department of Commerce
National Technical Information Service
NTIS

TECHNICAL REPORT DATA

(Please read Instructions on the reverse before completing)

1. REPORT NO. EPA 450/4-85-006	2.
3. RECIPIENT'S ACCESSION NO. PB85 242477/AS	
4. TITLE AND SUBTITLE Interim Procedures for Evaluating Air Quality Models: Experience with Implementation	5. REPORT DATE July 1985
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Monitoring and Data Analysis Division Office of Air Quality Planning and Standards U. S. Environmental Protection Agency Research Triangle Park, N.C. 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES	

16. ABSTRACT

This report summarizes and intercompares the details of five major regulatory cases for which guidance provided in the "Interim Procedures for Evaluating Air Quality Models" was implemented in evaluating candidate models. In two of the cases the evaluations have been completed and the appropriate model has been determined. In three cases the data base collection and/or the final analysis has not yet been completed. The purpose of the report is to provide potential users of the Interim Procedures with a description and analysis of several applications that have taken place. With this information in mind the user should be able to: (1) more effectively implement the procedures since some of the pitfalls experienced by the initial pioneers can now be avoided; and (2) design innovative technical criteria and statistical techniques that will advance the state of the science of model evaluation.

The analyses show that the basic principles or framework underlying the Interim Procedures is sound and workable in application. The concept of using the results from a prenegotiated protocol for the performance evaluation has been shown to be an appropriate and workable primary basis for objectively deciding on the best model. Similarly, "up-front" negotiation on what constitutes an acceptable data base network has been established as an acceptable way of promoting objectivity in the evaluation. Preliminary concentration estimates and the need for accurate continuous on-site measurements of the requisite model input data are also important.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Meteorology Mathematical Models Performance Evaluation Statistics	Performance Measures Technical Evaluation	4B 12A

18. DISTRIBUTION STATEMENT Unlimited	19. SECURITY CLASS (*This Report*) Unclassified	21. NO. OF PAGES
	20. SECURITY CLASS (*This page*) Unclassified	22. PRICE

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE

United States
Environmental Protection
Agency

Office of Air Quality
Planning and Standards
Research Triangle Park NC 27711

EPA-450/4-85-006
July 1985

Air

PB85-242477

EPA

Interim Procedures For Evaluating Air Quality Models: Experience with Implementation

REPRODUCED BY
NATIONAL TECHNICAL
INFORMATION SERVICE
U.S. DEPARTMENT OF COMMERCE
SPRINGFIELD, VA. 22161

EPA-450/4-85-006

Interim Procedures for Evaluating Air Quality Models: Experience with Implementation

U.S. ENVIRONMENTAL PROTECTION AGENCY
Monitoring and Data Analysis Division
Office of Air Quality Planning and Standards
Research Triangle Park, North Carolina 27711

July 1985

Disclaimer

This report has been reviewed by The Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency, and has been approved for publication. Mention of trade names or commercial products is not intended to constitute endorsement or recommendation for use.

Preface

In August 1981, EPA developed and distributed to its Regional Offices an in-house document "Interim Procedures for Evaluating Air Quality Models." The Regional Offices were encouraged to use the guidance contained in the document as an aid to determining whether a proposed model, not recommended in the Guideline on Air Quality Models[1], could be applied to a specific regulatory situation. Subsequently, as a result of experience gained in several applications of these procedures, EPA revised and published the "Interim Procedures for Evaluating Air Quality Models (Revised)"[2], in September, 1984.

The material contained in this report summarizes the experience gained from the first several applications of the original guidance. Potential users of the revised Interim Procedures are encouraged to read this report so that they might benefit from the experience of others and thus be able to better design their own application. The user should pay particular attention to the Findings and Recommendations (Section 4) so as to know and better understand particular aspects in the revised procedures on which EPA will place emphasis in the future applications.

Acknowledgements

This report was prepared by Dean Wilson with contributions from Joseph Tikvart, James Dicke and William Cox, all of the Source Receptor Analysis Branch, Monitoring and Data Analysis Division.

Appreciation is extended to Michael Koerber, Region V, Alan Cimorelli, Region III and Francis Gombar, Region II for their helpful comments during the review process. The patience of Linda Johnson as she typed this report is appreciated.

Table of Contents

		Page
Preface		iii
Acknowledgements		iv
Table of Contents		v
List of Tables		vii
List of Figures		ix
List of Symbols		xi
Summary		xiii
1.0 INTRODUCTION		1
1.1 Scope and Contents		2
1.2 Basic Principles Employed in the Interim Procedures		2
1.3 Summary of the Interim Procedures		3
2.0 APPLICATIONS OF THE INTERIM PROCEDURES TO REGULATORY PROBLEMS		7
2.1 Baldwin Power Plant		8
	2.1.1 Background	8
	2.1.2 Preliminary Analysis	10
	2.1.3 Protocol for the Performance Evaluation	11
	2.1.4 Data Bases for the Performance Evaluation	12
	2.1.5 Results of the Performance Evaluation and Model Acceptance	13
2.2 Westvaco Luke Mill		13
	2.2.1 Background	13
	2.2.2 Preliminary Analysis	14
	2.2.3 Protocol for the Performance Evaluation	15
	2.2.4 Data Bases for the Performance Evaluation	17
	2.2.5 Results of the Performance Evaluation and Model Acceptance	18
2.3 Warren Power Plant		19
	2.3.1 Background	19
	2.3.2 Preliminary Analysis	21
	2.3.3 Protocol for the Performance Evaluation	22
	2.3.4 Data Bases for the Performance Evaluation	25
2.4 Lovett Power Plant		25

2.4.1 Background 26
2.4.2 Preliminary Analysis 26
2.4.3 Protocol for the Performance Evaluation 28
2.4.4 Data Bases for the Performance Evaluation 30

2.5 Guayanilla Basin 30

2.5.1 Background 32
2.5.2 Preliminary Analysis 33
2.5.3 Protocol for the Performance Evaluation 34
2.5.4 Data Bases for the Performance Evaluation 37

2.6 Other Protocols 38

2.6.1 Example Problem 38
2.6.2 Gibson Power Plant 39
2.6.3 Homer City Area 40

3.0 INTERCOMPARISON OF APPLICATIONS 43

3.1 Preliminary Analysis 43

3.1.1 Regulatory Aspects 44
3.1.2 Source Characteristics and Source Environment 44
3.1.3 Proposed and Reference Models 46
3.1.4 Preliminary Concentration Estimates 48

3.2 Protocol for the Performance Evaluation 48

3.2.1 Performance Evaluation Objectives 49
3.2.2 Data Sets, Averaging Times and Pairing 50
3.2.3 Performance Measures 53
3.2.4 Model Performance Scoring 55

3.3 Data Bases for the Performance Evaluation 57

3.4 Negotiation of the Procedures to be Followed 60

4.0 FINDINGS AND CONCLUSIONS 63

5.0 REFERENCES 69

Appendix A. Protocol and Performance Evaluation Results for Baldwin Power Plant A-1

Appendix B. Protocol and Performance Evaluation Results for Westvaco Luke Mill B-1

Appendix C. Protocol for Warren Power Plant C-1

Appendix D. Protocol for Lovett Power Plant D-1

Appendix E. Protocol for Guayanilla Basin E-1

List of Tables

Number	Title	Page
3-1	Source and Source Environment	45
3-2	Proposed and Reference Models	47
3-3	Weighting of Maximum Possible Points by Data Set, Averaging Time and Degree of Pairing	51
3-4	Performance Measures Used in the Protocols	54
3-5	Data Bases for Performance Evaluations	58
3-6	Issues Involved in Negotiations	62

List of Figures

Number	Title	Page
1-1	Decision flow diagram for evaluating a proposed air quality model	4
2-1	Map of air quality monitoring stations and the meteorological tower in the vicinity of the Baldwin power plant, April 1982-March 1983	9
2-2	Topographic map of the area surrounding the Westvaco Luke Mill	14
2-3	Map of seven air quality monitoring stations and the meteorological station in the Warren area	20
2-4	Map of air quality monitoring stations and the primary meteorological tower in the vicinity of the Lovett power plant	26
2-5	Map of existing air quality monitoring network and expanded air quality monitoring network in the Guayanilla area	31

List of Symbols

C_o = Observed Concentration

C_p = Predicted Concentration

d = Residual = $C_o - C_p$

M_c = Number of Observed/Predicted Meteorological Events in Common

R = Pearson's Correlation Coefficient

$RMSE_d$ = Root-mean-square-error of Residual

S_d = Standard Deviation of Residual

S_o^2 = Variance of Observed Concentration

S_p^2 = Variance of Predicted Concentration

Summary

This report summarizes and intercompares the details of five major regulatory cases for which the guidance provided in the "Interim Procedures for Evaluating Air Quality Models"* was implemented in evaluating candidate models. In two of the cases the evaluations have been completed and the appropriate model has been determined. In three cases the data base collection and/or the final analysis has not yet been completed.

Due to the unique source-receptor relationships in each case, however, the procedures, data bases and number of monitors here are not necessarily applicable to other situations. These cases are presented only as examples of how the 1981 Interim Procedures document has been applied to some real world situations.

Each of the five cases involves major point sources of SO_2. In all cases the major regulatory concern is to determine the emission limit that would result in attainment of the National Ambient Air Quality Standards (NAAQS) within a few kilometers of the plants. Most of the cases involve power plants and/or industrial facilites located in complex terrain where short-term impact on nearby terrain is the critical source-receptor relationship.

Although the scope of model problems is limited, it seems clear that the basic principles or framework underlying this guidance is sound and workable in application. For example, the concept of using the results from a pre-negotiated protocol for the performance evaluation has been shown to be an appropriate and workable primary basis for objectively deciding on the best model. Similarly, "up-front" negotiation on what constitutes an acceptable data base network, while often difficult to accomplish because of conflicting

---

*1981 EPA internal document

viewpoints, has been established as an acceptable way of promoting objectivity in the evaluation.

In earlier evaluations there was some laxity on the part of the reviewing agencies in requiring a detailed preliminary evaluation/documentation of the critical source-receptor relationships. In more recent evaluations fulfilling the requirement for preliminary estimates has led to better understanding of the source-receptor relationships and provided a better linkage between these relationships and the contents of the performance evaluation protocol. These preliminary estimates also seem to better define the requisite data base network. As a consequence of this experience, it is recommended that in future protocols more emphasis be placed on the preliminary analysis; the results of this analysis should be linked to the protocol and the requisite data base through the development of detailed performance evaluation objectives.

Experience has also pointed up the need to build in some "safeguards" in the application of the chosen model, should that model be shown to underpredict concentrations. This is particularly a problem if an emission limit derived from the model application might result in violations of the NAAQS. The methods used in more recent regulatory cases generally involve the use of "adjustment factors" to correct for possible underprediction. This technique is not particularly appealing and the development of more innovative and scientifically defensible schemes is recommended.

Finally, based on this experience, it should be emphasized that the credibility of the performance evaluation is greatly enhanced by the availability of continuous on-site measurements of the requisite model input data. This includes the measurement of meteorological parameters, as well as pre-specified backup data sources for missing data periods. Also included is the need for continuous in-stack measurement of emissions and accurate stack parameter data.

1.0 INTRODUCTION

In 1981 a document "Interim Procedures for Evaluating Air Quality Models" was prepared in-house by EPA and distributed to the ten Regional Offices. This document identified the documentation, model evaluation and data analyses desirable for establishing the appropriateness of a proposed model. The Regional Offices were encouraged to use the procedures when judging whether a model not specifically recommended for use in the "Guideline on Air Quality Models,"[1] was acceptable for a given regulatory action. These procedures, which involved the quantitative evaluation and comparison of models for application to specific air pollution problems, addressed a relatively new problem area for the modeling community. It was recognized that experience with their use would provide better insight to the model evaluation problem and its limitations. During the 1981-1984 time period, several projects which entailed the use of the procedures were undertaken. Based on this experience, the procedures were revised and published as "Interim Procedures for Evaluating Air Quality Models (Revised)"[2].

It was clear from the experience gained in application of these 1981 procedures that the basic principles contained therein were sound and appropriate to apply to regulatory model evaluation problems. However the state of the science did not suggest a single prescription detailing their application. In fact, each application of the procedures differed considerably in detail. However, while the individual merits of each application could be scientifically debated, each case reflected an acceptable interpretation of the interim guidance.

1.1 Scope and Contents

The purpose of this document is to provide potential users of the revised Interim Procedures with a description and analysis of several applications that have taken place. With this information in mind the user should be able to: (1) more effectively implement the procedures since some of the pitfalls experienced by the initial pioneers can now be avoided; and (2) design innovative technical criteria and statistical techniques that will advance the state of the science of model evaluation.

Remaining sections of this report are as follows. Section 1.2 reviews the basic principles underlying the Interim Procedures. Section 1.3 is a summary of the Interim Procedures, to be used as a point of reference in reading this report. Section 2 contains summaries of each of five major regulatory cases where the Interim Procedures were applied, as well as brief summaries of three other incomplete cases. Section 3 intercompares the technical details of each of the five cases. Section 4 lists the findings and recommendations resulting from the analyses in Sections 2 and 3. Appendices A-E contain details of the protocols for each of the five cases. Appendices A and B also contain the final scores for two of the performance evaluations.

1.2 Basic Principles Employed in the Interim Procedures

The Interim Procedures for Evaluating Air Quality Models is built around a framework of basic principles whereby the details of the decision process to be used in the model evaluation should be established and documented up-front. The performance evaluation protocol should be established before data are available that would allow either the applicant or the control agency(s) to determine, in advance, the outcome of the evaluation. These principles are:

° Up-front negotiations/agreements between the user and the regulatory agencies are vital;

° All relevant technical data/analyses and regulatory constraints are documented;

° A protocol for performance evaluation is written before any data bases are in hand;

° A data base network is established that will meet the needs of both the technical/regulatory requirements and the performance evaluation protocol;

° The performance evaluation is carried out and the decision on the appropriate model must be made as prescribed in the protocol.

The material in Sections 2 and 3 is an analysis, among other things, of how well these principles were adhered to for five cases. The findings in Section 4 include specific statements to this effect.

1.3 Summary of the Interim Procedures

The document Interim Procedures for Evaluating Air Quality Models (Revised) describes procedures for use in accepting, for a specific application, a model that is not recommended in the Guideline on Air Quality Models. One requirement is for an evaluation of model performance. The primary basis for the model evaluation assumes the existence of a reference model which has some pre-existing status and to which the proposed nonguideline model can be compared from a number of perspectives. However for some applications it may not be possible to identify an appropriate reference model, in which case specific requirements for model acceptance must be identified. Figure 1-1 provides an outline of the procedures described in the document.

After analysis of the intended application, or the problem to be modeled, a decision must be made on the reference model to which the proposed

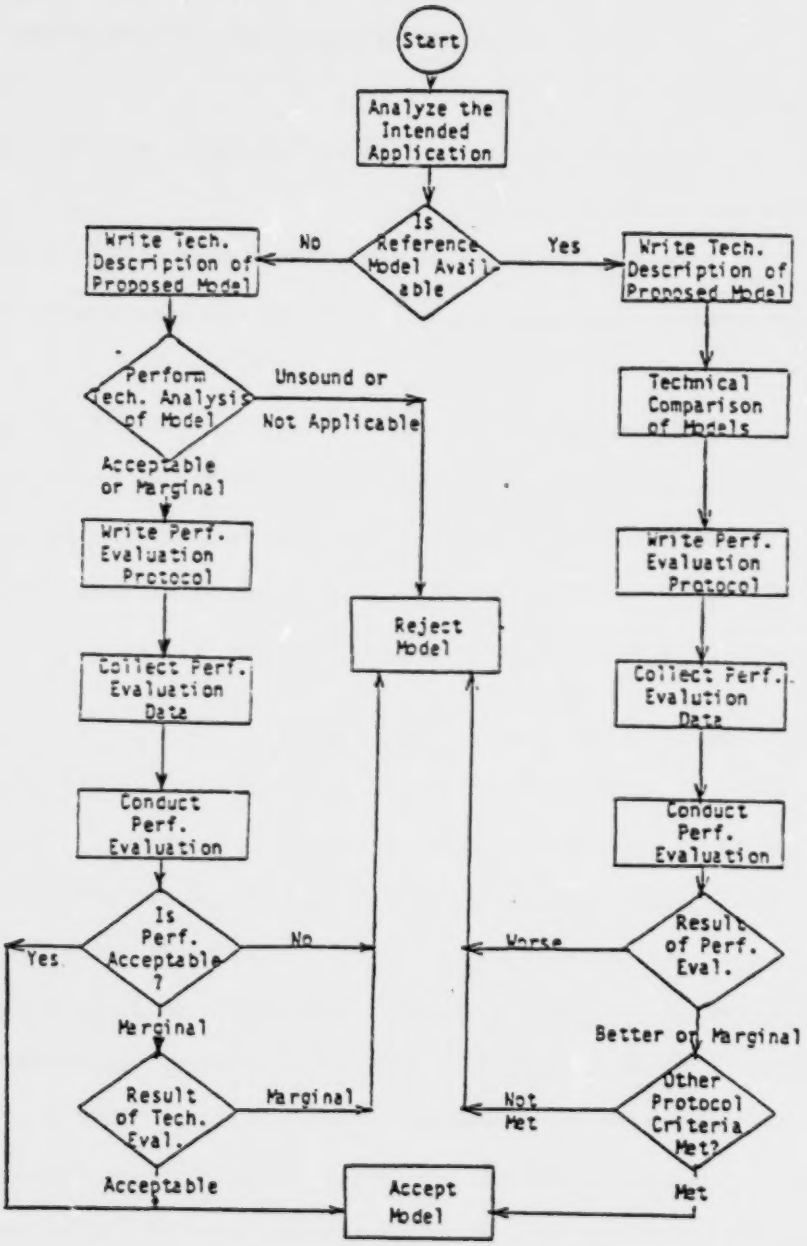


Figure 1-1. Decision flow diagram for evaluating a proposed air quality model

model can be compared. If an appropriate reference model can be identified, then the relative acceptability of the two models is determined as follows. The model is first compared on a technical basis to the reference model to determine if it can be expected to more accurately estimate the true concentrations. This technical comparison should include preliminary concentration estimates with both models for the intended application. Next a protocol for model performance comparison is written and agreed to by the applicant and the appropriate regulatory agency. This protocol describes how an appropriate set of field data will be used to judge the relative performance of the proposed and the reference model. Performance measures recommended by the American Meteorological Society[3] are used in describing the comparative performance of the two models in an objective scheme. That scheme should consider the relative importance to the problem of various modeling objectives and the degree to which the individual performance measures support those objectives. Once the plan for performance evaluation is written and the data to be used are collected/assembled, the performance measure statistics are calculated and the weighting scheme described in the protocol is executed. Execution of the decision scheme will lead to a determination that the proposed model performs better, worse or about the same as the reference model for the given application. The final determination of the acceptability of the proposed model should be based primarily on the outcome of the comparative performance evaluation. However, if so specified in the protocol, the decision may also be based on results of the technical evaluation, the ability of the proposed model to meet minimum standards of performance, and/or other specified criteria.

If no appropriate reference model is identified, the proposed model is evaluated as follows. First the proposed model is evaluated from a technical standpoint to determine if it is well founded in theory, and is applicable to the situation. Preliminary concentration estimates for the proposed application should be included. This involves a careful analysis of the model features and use in comparison with the source configuration, terrain and other aspects of the intended application. Secondly, if the model is considered applicable to the problem, it is examined to see if the basic formulations and assumptions are sound and appropriate to the problem. (If the model is clearly not applicable or cannot be technically supported, it is recommended that no further evaluation of the model be conducted and that the exercise be terminated.) Next, a performance evaluation protocol is prepared that specifies what data collection and performance criteria will be used in determining whether the model is acceptable or unacceptable. Finally, results from the performance evaluation should be considered together with the results of the technical evaluation to determine acceptability.

2.0 APPLICATION OF THE INTERIM PROCEDURES TO REGULATORY PROBLEMS

This section describes five major regulatory cases, covering the period 1982-1984, where the techniques described in the Interim Procedures are being applied to establish the appropriate model for setting emission limits. Although protocols for the comparative performance evaluation of competing models have been prepared for all five cases, in only two cases has the execution of the protocol been completed; these results are presented.

Sections 2.1 through 2.5 are arranged roughly chronologically, i.e. in the order in time when a final performance evaluation protocol was established. Section 2.6 contains brief summaries for other applications of the Interim Procedures of which EPA is aware; however, for a variety of reasons, the chosen models have not been used in regulatory decision-making.

The history of negotiation over appropriate models, data bases, emission limits, etc. for the sources included in these specific applications dates back several years. The development and execution of an agreed upon procedure for the comparative performance evaluation of competing models is, or is designed to be, the basis for resolution of these issues. No attempt is made in the following subsections to describe the complete history of issues/negotiations. Instead, only a brief definition of the issues to be resolved by the performance evaluation is provided.

Each of the Sections 2.1 through 2.5 contain separate subsections dealing with the background (history), the preliminary analysis, the protocol for the performance evaluation and the data bases to be used in the performance evaluation. In addition, Sections 2.1 and 2.2 include a subsection which summarizes the results of the performance evaluation.

2.1 Baldwin Power Plant

The Baldwin power plant, located in Randolph County, Illinois, about 60 km southeast of St. Louis, Missouri, is composed of three steam/electric generating units with a combined design generating capacity of 1,826 megawatts. Each of the boilers is vented through an individual 605-foot (184m) stack. A map of the area is provided in Figure 2-1.

2.1.1 Background

In late 1981 the State-approved SO_2 emission rate was 101,588 lb/hour. Illinois Power Company (IP) requested that this rate be established as the EPA-approved SIP limit adequate to protect both primary and secondary National Ambient Air Quality Standards (NAAQS). The basis for this proposal was estimates by the MPSDM model indicating compliance with the standards. The company claimed that the use of MPSDM was supported by data from an 11-station monitoring network in the vicinity of the plant. Estimates using the EPA CRSTER model indicated compliance with the primary NAAQS but violations of the 3-hour secondary NAAQS.

Potential problems were:

1. locations of the monitors were not adequate to conduct a performance evaluation for MPSDM and CRSTER.

2. adequacy of the IP model performance evaluation was in question since the available monitoring data were used to select a "best fit" option of MPSDM, i.e., an independent performance evaluation was not conducted; and

3. available monitoring data (summarized as block data) indicated exceedances (no violations) of both the 3-hour and 24-hour standards at a previously operated monitor not included in the 11-station network.

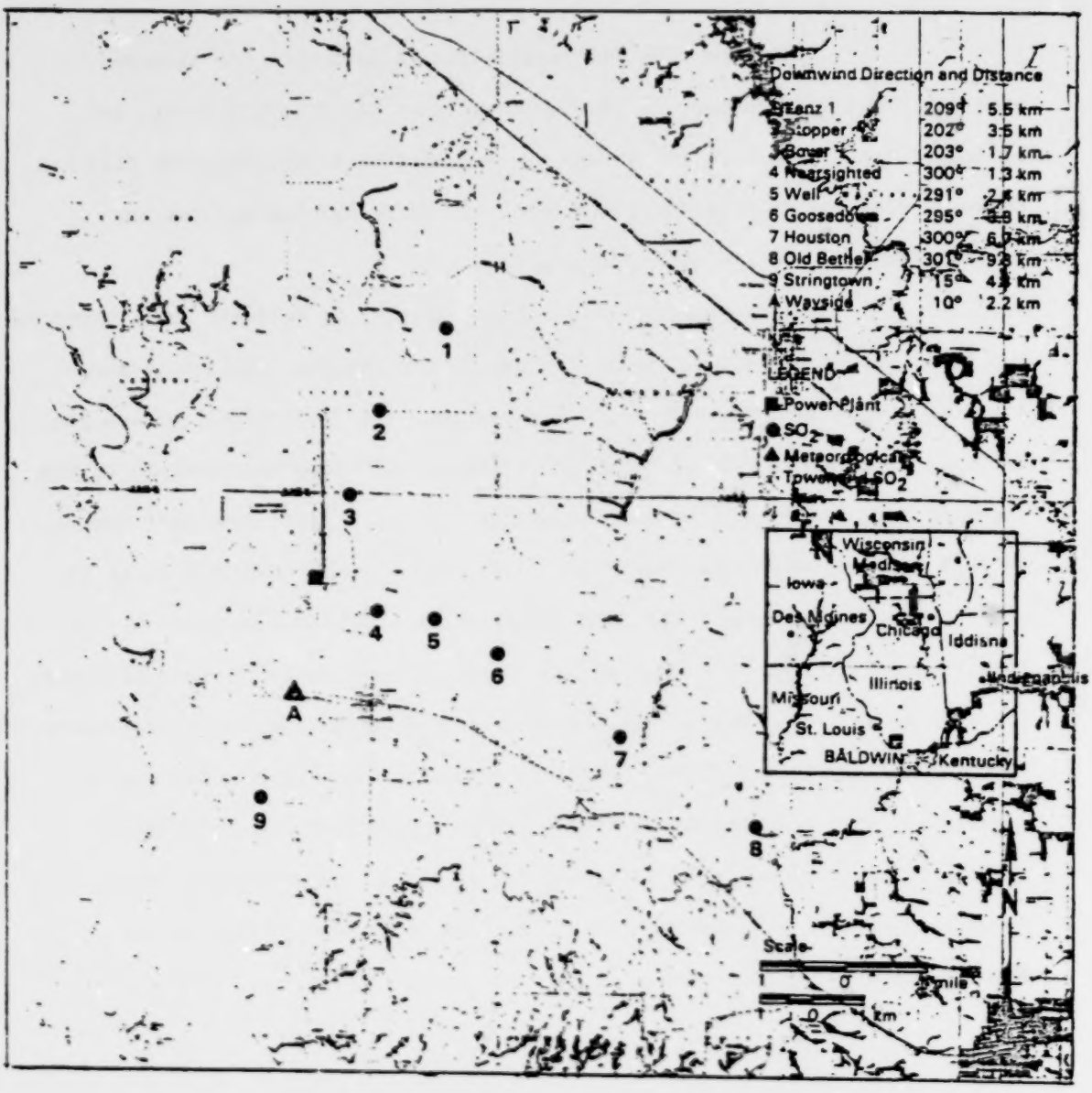


Figure 2-1. Map of air quality monitoring stations and the meteorological tower in the vicinity of the Baldwin power plant, April 1982-March 1983.

Based on this information EPA decided that the proposed emission limit was adequate to attain the primary SO_2 NAAQS; however, the secondary NAAQS demonstration should be re-evaluated by the State of Illinois. Guidance contained in the Interim Procedures for Evaluating Air Quality Models should be used in the re-evaluation.

In response to this suggestion, IP, in February 1982, prepared the "Proposed Procedures for Model Evaluation and Emission Limit Determination for the Baldwin Power Plant." Negotiations then took place between the Illinois Environmental Protection Agency (IEPA) and IP on the contents of the document. The end result of these negotiations was a final protocol[4] issued by IEPA in June 1982. The four major differences between the IEPA document and the IP protocol were: (1) IEPA eliminated one performance measure that involved case studies of the 10 episodes with highest measured concentrations, (2) more weight was given to the comparison of the second-high, single-valued residuals in the IEPA protocol (and less weight for some of the other measures); (3) IEPA eliminated the use of 1-hour statistics; and (4) IEPA eliminated performance tests involving comparison of monitored data with predictions for a 180 receptor grid. (Instead, only predictions at the monitor sites were to be used.)

2.1.2 Preliminary Analysis

The preliminary analysis of the proposed application, submitted by IP to IEPA, included a definition of the regulatory aspects of the problem and a description of the source and its surroundings. The analysis established that only the 3-hour concentration estimates were at issue. IP proposed to use MPSDM in lieu of CRSTER to estimate 3-hour concentrations, pending the outcome of a comparative performance evaluation. A technical description of MSPDM and a user's manual for the model were

provided to IEPA. IP also provided a technical comparison between MPSDM and CRSTER following the procedures outlined in the "Workbook for Comparison of Air Quality Models"[5]. IP's "workbook" comparison concluded that MPSDM was technically comparable to CRSTER for most application elements but was technically better for two of the elements; thus MPSDM was judged by IP to be technically superior to CRSTER for the proposed application.

Preliminary concentration estimates were made with both CRSTER and MPSDM although the details of these estimates were not documented. From other information available it was evident that MPSDM would yield lower 3-hour estimates than CRSTER at locations within 2 km under very unstable meteorological conditions (A-stability). These estimates would be controlling, i.e. the estimates that would be used to set the emission limit for the power plant.

2.1.3 Protocol for the Performance Evaluation

The IEPA protocol for the comparative performance evaluation of MPSDM and CRSTER, which is detailed in Appendix A, strongly emphasized accurate prediction of the peak (highest-second-highest) estimate. Fifty-five (55) percent of the weighting in the protocol involved the calculation of performance statistics that characterize each model's ability to reproduce the measured second-high concentrations at the various monitors. Thirty-five (35) percent of the weighting was assigned to performance statistics that characterize the models' ability to reproduce the measured concentration in the upper end of the observed frequency distribution, namely the high-25 observed and predicted concentrations. In addition, the protocol included performance measures designed to determine how well the models perform for specific meteorological conditions (5%) and performance statistics that compare the upper end of the frequency distribution of measured/predicted values (5%).

The primary performance measures used in the evaluation were the residual (observed minus predicted concentration) and the bias (average residual for the high-25 data set). Performance measures were calculated from data paired in space and time and completely unpaired with the major weighting on the unpaired data. Other performance measures in the protocol included the standard deviation of the residual and the root-mean-square-error of the residual.

The scoring scheme used for most performance statistics consisted of a percentage of maximum possible points within specified cutoff values. If the performance statistic fell outside of the cutoff values, no points were awarded to the model. Within the acceptable range, the percent of possible points was linearly related to the value of the performance statistic. The sign (+ or -) of the residual and bias statistics was not considered in the scoring process, i.e. overprediction and underprediction were weighted equally. The scoring schemes for the meteorological cases and for the frequency distrubutions were more complicated; refer to Appendix A for details.

The decision criteria by which the better model was chosen was simply which model attained the best score.

2.1.4 Data Bases for the Performance Evaluation

As mentioned earlier, the data base for the performance evaluation ultimately consisted of a network of monitors and a meteorological station specifically designed to fit the needs of the application (See Figure 2-1). Data obtained from previously operated networks were used in designing this data base network. This data base consisted of 10 SO_2 monitors and a single meteorological tower instrumented to collect wind speed, wind direction and turbulence intensity (for use in MPSDM) data. Off-site meteorological

data used in the evaluation consisted of mixing height data derived from National Weather Service (NWS) soundings from Salem, IL and Pasquill-Gifford stability data derived from surface observations at Scott Air Force Base, IL (CRSTER only). Hourly emission data and stack gas parameters were derived from records of plant load level and daily coal samples.

2.1.5 Results of the Performance Evaluation and Model Acceptance

The data base for this evaluation has been collected and the performance evaluation has been carried out according to terms specified in the protocol[6]. The overall result was that MPSDM scored 51.3 points and CRSTER scored 41.7 points out of a possible 100 points. Thus MPSDM was selected as the appropriate model to be used to determine the emission limit necessary to attain the secondary 3-hour NAAQS. Details of the performance evaluation results are provided in Appendix A.

2.2 Westvaco Luke Mill

The Westvaco Luke mill in the town of Luke in Allegany County, Maryland, is located 970 feet (296m) above mean sea level (msl) in a deep valley on the north branch of the Potomac River. The region surrounding the mill is mountainous and generally forested. Figure 2-2 is a topographic map of the area surrounding the Westvaco Luke Mill. The ⊙ symbol shows the location of the 623-foot (190m) main stack which serves the facility.

2.2.1 Background

In response to a consent decree, the company operated an ambient monitoring and meteorological data collection network from December 1979 through November 1981. The ■ symbols in Figure 2-2 show the location of continuous SO_2 monitors and the ▲ symbols show the locations of the

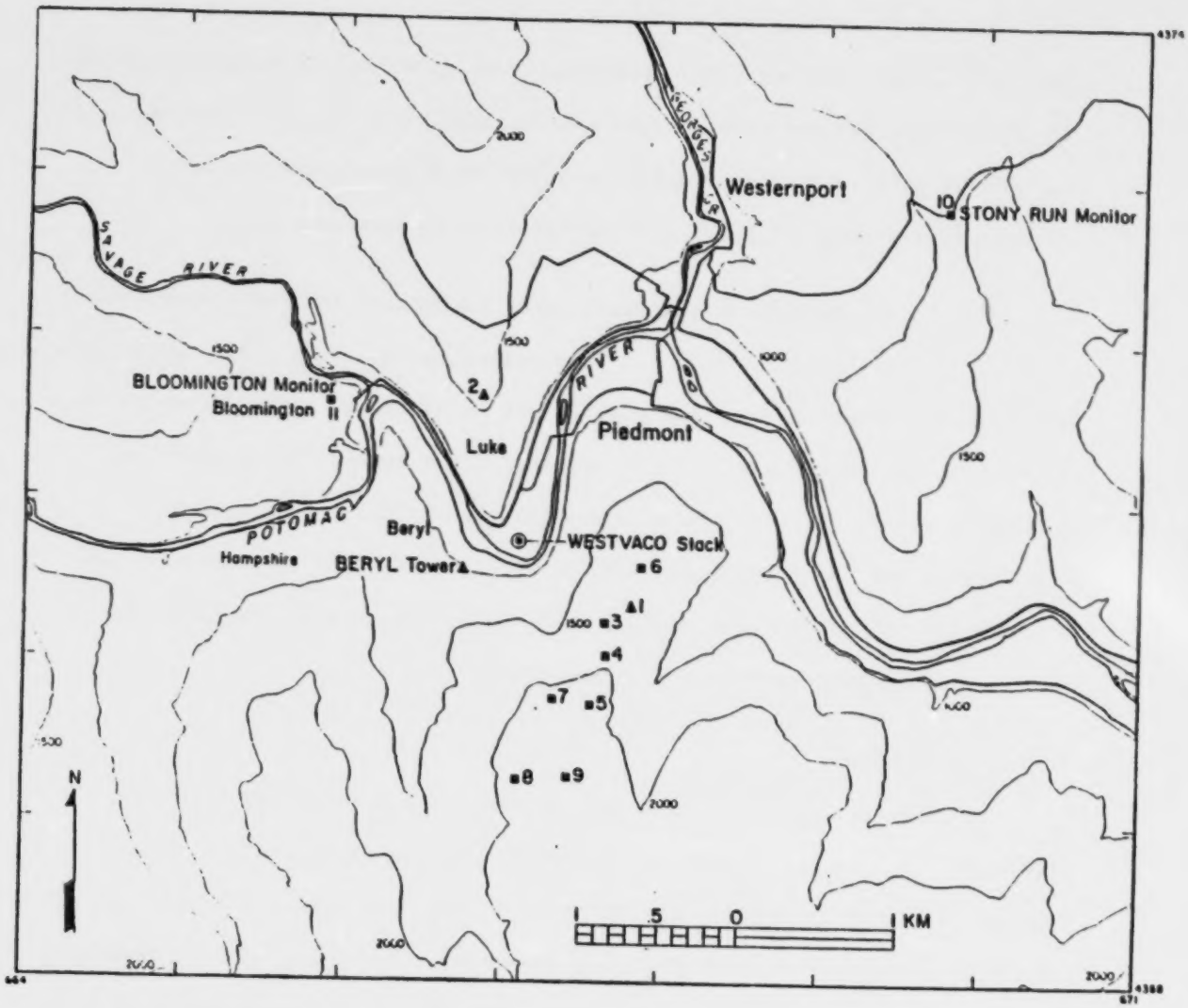


Figure 2-2. Topographic map of the area surrounding the Westvaco Luke Mill. Elevations are in feet above mean sea level and the contour interval is 500 feet. The ■ symbols represent SO_2 monitoring sites. The ▲ symbols represent meteorological monitoring sites. Sites 1 and 2 are also SO_2 monitoring sites.

100-meter Meteorological Tower No. 1, the 30-meter Meteorological Tower No. 2 (the Luke Hill Tower) and the 100-meter Beryl Meteorological Tower. Continuous SO_2 monitors were collocated with Tower No. 1 and Tower No. 2 and an acoustic sounder was collocated with Tower No. 2. As shown by Figure 2-2, there were eleven SO_2 monitors of which eight were located on a ridge southeast of the Main Stack. SO_2 emissions during the two-year monitoring period were limited to 49 tons per day.

The company developed a site-specific dispersion model, LUMM, which they claimed was applicable to the problem and should be accepted as the basis for setting a new emission limit of 89 tons per day. The company's basis for this claim was described in a March 1982 report[7] in which estimates from the LUMM model were compared to ambient measurements from the 11-station network. EPA reviewed the report and found a number of technical problems with the model, including the use of ambient data to "tune" the model, i.e. no independent performance evaluation was undertaken.

In order to resolve these problems, EPA developed, under contract in mid-1982, a protocol for conducting a performance evaluation of models applicable to the Westvaco site. The company was then asked to compare their model with the SHORTZ model, using procedures like those suggested in this protocol. As a result of these negotiations a final protocol[8] was agreed upon in late 1982 and subsequently executed by the company, utilizing the second year of the two-year data base.

2.2.2 Preliminary Analysis

There is little written material on the Westvaco case which would suggest that an up-front, in-depth preliminary analysis of regulatory and technical aspects of the problem was undertaken. However, based on the

above two references, various Federal Register actions and numerous meetings, both the source and the control agencies apparently had at least tacit understandings of the regulatory and technical issues involved. For example, the regulatory agencies were concerned about attainment of the short-term ambient standards at elevated receptors near (within a few kilometers of) the source. It was also apparent that SHORTZ would yield higher concentration estimates, and thus a tighter emission limit, than LUMM.

References 7 and 8 contain technical descriptions of the two competing models but no user's manuals. The SHORTZ model was modified for use at Westvaco and no user's manual exists for this version. The references do not describe any preliminary estimates using the two models nor do they contain an in-depth technical comparison of the two models. No analysis using the Workbook for Comparison of Air Quality Models was undertaken.

2.2.3 Protocol for the Performance Evaluation

The final agreed upon protocol for the comparative performance evaluation of LUMM and SHORTZ, which is detailed in Appendix B, emphasized accurate estimates of the peak concentrations and the upper end of the frequency distributions. Forty-three (43) percent of the weighting in the protocol involved the calculation of performance statistics that characterize each model's ability to reproduce the measured maximum and second-high concentrations at the various monitors. Fifty-seven (57) percent of the weighting was assigned to performance statistics that characterize the models' ability to reproduce measured concentrations in the upper end of the observed frequency distribution, namely the high-25 observed and predicted concentrations. No "all data" statistics were calculated, i.e. the protocol assumed that the only relevant data were the top-25 estimated and observed concentrations.

The protocol specified three basic performance measures to be used in the evaluation, the absolute residual for single-valued comparisons, the bias for the top-25 concentrations and the ratios of the observed and predicted variances for the the top-25 concentrations. Various time and space pairings were specified with most of the weighting (61 percent) on data paired in space but not time.

The scoring scheme used for each performance statistic was specified by somewhat complicated formulae and the reader is referred to Appendix B for details. Basically, the scheme involved computing ratios of performance measures between the two competing models and bias ratios or variance ratios for each model. These ratios were then combined in various ways to produce a percentage of maximum possible points for each performance statistic. This result was then multiplied by the maximum possible points for that performance statistic to yield a subscore. Subscores were then totalled for each model to yield a composite score. The model with the highest total score was deemed to be most appropriate to apply to the source.

2.2.4 Data Bases for the Performance Evaluation

The data base used in the performance evaluation was the second year of the historical two-year data base described above. The locations of the ten monitors and the two meteorological towers for use in the evaluation are shown in Figure 2-2. Data from the Beryl tower and the Bloomington monitor were not to be used, although Bloomington data were used to help establish background values. Each tower was instrumented at a number of levels; thus there were often a number of possible values for the meteorological inputs to each model to choose from. To promote objectivity in the evaluation, the primary source of data for each meteorological

parameter, as well as ranked "default" data sources to be used in the event of missing data, were specified in a protocol. No off-site meteorological data were used in the models; however default values for mixing height and some turbulence intensities were specified. Hourly emission data and stack gas parameters were derived from continuous in-stack measurements.

The data base for the model evaluation already existed. The network was designed in 1978-1979, to determine if there were any NAAQS violations in the vicinity of the plant and possibly for use in conducting a performance evaluation. Most of the monitors were densely clustered on the hillside south of the plant, the area where maximum concentrations were expected. However, a decision was made that the definition of ambient air did not apply to this property, i.e. the NAAQS did not apply there. This fact, together with an opinion of the control agencies that the LUMM model was partially based on the same data that would be used in the performance evaluation, raised many questions on the objectivity of the evaluation. Detailed records on the negotiations between the company and control agencies to resolve this concern are lacking. In the end it was apparently decided that the objectivity in the performance evaluation was not sufficiently compromised to warrant the redesign of the network and collection of an additional year's data. The performance evaluation protocol contained one mitigating measure in this regard. In an apparent attempt to compensate for the lack of sufficient "offsite" monitors, several of the performance statistics for the single offsite monitor (No. 10 in Figure 2-2) were weighted by a factor of four over those same statistics for the other eight monitors.

2.2.5 Results of the Performance Evaluation and Model Acceptance

The data base for this evaluation has been collected and the performance evaluation has been carried out according to terms specified

in the protocol[9]. The overall result was that LUMM scored 363 points and SHORTZ scored 168 points out of a possible 602 points. Thus, LUMM was selected as the appropriate model to be used to determine the emission limit necessary to attain the NAAQS. Details of the performance evaluation results are provided in Appendix B.

2.3 Warren Power Plant

The 90 megawatt Warren power plant, operated by the Pennsylvania Electric Company (Penelec), is located in Warren County in northern Pennsylvania, about 80 km southeast of Erie. The plant has a single 200-foot (61m) stack which emits about 2420 lb/hour of SO_2 at maximum capacity. The modeling region near Warren is characterized by irregular mountainous terrain, with peak terrain elevations substantially above the top of the power plant stack (See Figure 2-3).

2.3.1 Background

As a result of earlier modeling, the area was designated as nonattainment in the late 1970's. Penelec was directed by the State of Pennsylvania to establish, through monitoring and modeling, an emission limit that would ensure attainment of the NAAQS.

Penelec believed that the LAPPES model was appropriate to use for purposes of setting the emission limits. In March 1984 Penelec proposed to the State of Pennsylvania Department of Environmental Resources (DER) an analysis and a performance evaluation protocol, patterned after the Interim Procedures, to establish whether LAPPES would be more appropriate than EPA's Complex I model. A series of negotiations between DER, EPA and Penelec followed. A number of additions and changes were made to

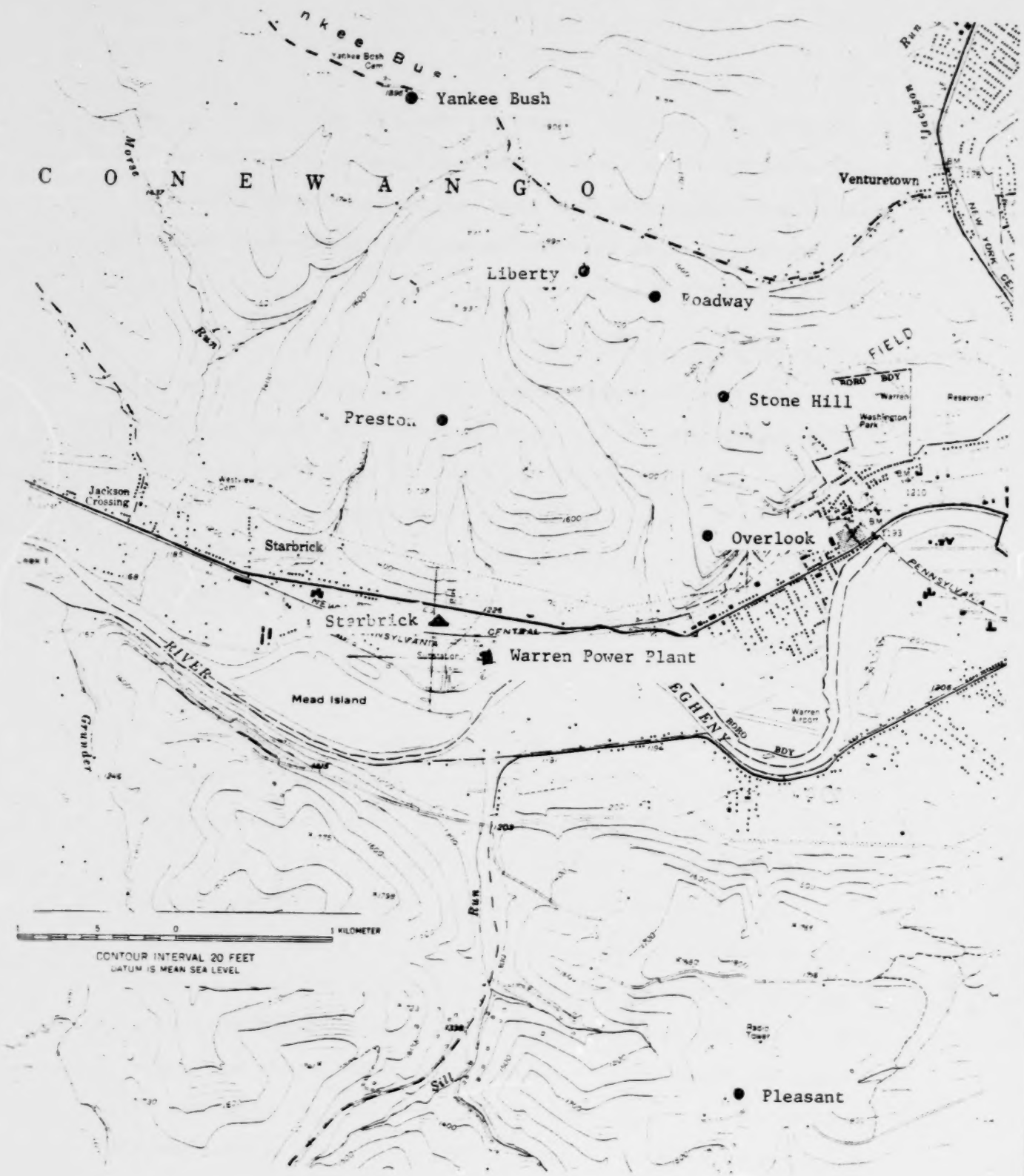


Figure 2-3. Map of seven air quality monitoring stations (●) and the meteorological station (▲) in the Warren area.

the analysis and the protocol and a final agreed upon analysis and protocol was written in November 1984[10]. Data collection requisite to executing the protocol is currently underway.

2.3.2 Preliminary Analysis

The protocol document contains a definition of the regulatory aspects of the problem and a description of the source and surroundings. The analysis establishes that the 3-hour and the 24-hour concentration estimates are at issue. Penelec proposes to use LAPPES in lieu of Complex I to estimate concentrations for all averaging times pending the outcome of a comparative performance evaluation. Penelec has also submitted a technical description of LAPPES. Although a user's manual for LAPPES exists, it is not clear that the manual is "current" with the version of LAPPES used in this application. Penelec has not provided a rigorous technical comparison of LAPPES and Complex I following the procedures outlined in the Workbook for Comparison of Air Quality Models.

Based on one year of meteorological data, preliminary concentration estimates have been made with both LAPPES and Complex I and the details of these estimates, including isopleth maps, are provided in the protocol document. These estimates show that maximum concentrations for all averaging times occur on elevated terrain to the north of the plant. The preliminary analysis also identifies another significant SO_2 source located approximately 4 km east of the Warren power plant. This source is close enough such that short-term impacts could overlap. Since monitoring data would not always distinguish between these sources, both sources are included in the model comparison study. The hourly average background SO_2 concentration is to be the lowest concentration observed by any station in the monitoring network.

2.3.3 Protocol for the Performance Evaluation

The protocol for the comparative performance evaluation of LAPPES and Complex I, which is detailed in Appendix C, emphasizes accurate prediction of the peak concentration. Forty-three (43) percent of the weighting in the protocol involves the calculation of performance statistics that characterize each model's ability to reproduce the measured high and second-high concentrations at the various monitors. An additional forty-three (43) percent of the weighting is assigned to performance statistics that characterize the models' ability to reproduce the measured concentration in the upper end of the observed frequency distribution, namely the high-25 concentrations. These analyses of the high-25 data set include certain statistics that break out performance by stability category. In addition, the protocol assigns a weight of fourteen (14) percent to performance statistics on the entire range (all data) of measured/predicted values.

A variety of performance measures are used in the Warren protocol; see Appendix C. Although the bias is weighted heavily in all of the data sets, the specific performance measures used to characterize bias vary. For the maximum single-valued comparisons, the average residual and the ratio of the absolute residual to the observed concentration (both paired in space but not time) are used to characterize the bias. For other data sets, including the second high single-valued comparisons, extensive use is made of the ratio of the predicted to observed concentrations as a measure of bias. Other performance measures used in the protocol include correlation measures and ratios of predicted to observed variances. Performance statistics are to be calculated for 1-hour, 3-hour, 24-hour and annual averaging times. Each averaging time carries considerable weighting. Sixty (60) percent of the weighting is assigned to unpaired data comparisons and forty (40) percent to data paired in space but not time.

The scoring scheme used for most performance statistics consists of a percentage of maximum possible points within specified cutoff values. If the performance statistics fall outside of the cutoff values, no points are to be awarded to the model. Within the acceptable range, the percent of possible points is specified in tabular form (discrete values for specified ranges of performance). The tabular values for the bias statistics slightly favor the model that overpredicts, if one model overpredicts to the same extent that the other model underpredicts. The scoring schemes for other performance measures are more complicated; refer to Appendix C for details. Subscores for each performance statistic are totaled to obtain a final score for each model.

Initially, the model with the highest score is deemed to be most appropriate to apply for regulatory purposes. However, the protocol contains some additional procedures to be employed if the LAPPES model attains the highest score but is shown to underpredict the highest concentrations. For the 3-hour and 24-hour averaging periods, the average of the 10 highest concentrations predicted by LAPPES will be compared with the average of the 10 highest observed values. If the ratio of the observed to predicted average is greater than 1, then this ratio will be used to adjust LAPPES model predictions for the regulatory analyses. This "safety factor" is intended to compensate for any systematic model underpredictions. If the ratio is less than 1, no adjustment will be made. Note that a different ratio will be used for each averaging time. For annual average concentrations, the averages of observed and predicted values at the seven monitoring stations will be compared. If the average of observed annual values is larger than predicted, then model predictions will be adjusted by the ratio of the observed to predicted average.

2.3.4 Data Bases for the Performance Evaluation

The data base for the performance evaluation consists of a network of monitors and meteorological stations specifically designed to cover the area of maximum predicted concentration and to fit the needs of the protocol. This data base consists of seven monitors, six of which are in the area north of the plant, where preliminary estimates indicated that high concentrations would occur (See Figure 2-3). The seventh monitor, located south of the plant would most often be used to determine background. Two meteorological towers are included in the network but data from the Starbrick tower would be used exclusively unless such data are missing. For missing data periods, a hierarchy of default data sources are specified in the protocol, including data from the Preston tower and off-site data. Wind fluctuation (sigma theta) data are used to determine stability in accordance with the scheme defined in the "Regional Workshops on Air Quality Modeling: A Summary Report"[11]. Morning and afternoon mixing heights are primarily from Pittsburgh National Weather Service data. Hourly emission data and stack gas parameters are to be derived from records of plant load level and coal sample data.

2.4 Lovett Power Plant

The Lovett power plant is located in the Hudson River Valley of New York State and is owned by Orange and Rockland Utilities, Inc. The plant generates 495 megawatts of electricity and is currently burning 0.37 percent sulfur oil. Major terrain features in the vicinity of the plant include the Hudson River Valley, which generally runs from north to south, and several nearby mountains. Dunderberg Mountain, with a maximum elevation of approximately 1100 feet (335m) is located 1-2 km to the north. Other significant topographic elevations include Buckberg Mountain, about 1.3 km to the

west, with a peak of 787 feet (240 m). An area of high terrain extends from west-northwest through north within 5 km of the plant. A map of the region is presented in Figure 2-4.

2.4.1 Background

The company requested to convert the plant to low sulfur (0.6-0.7 percent) coal with a new emission limit of 1.0 lbs SO_2/mm Btu. An actual increase in SO_2 emissions of approximately 12,000 tons per year would result.

In April 1984, the EPA Administrator agreed, in principle, to allow the company to construct a new 475-foot (145m) stack and convert the plant to coal. One provision of the agreement was that the company develop a protocol for a performance evaluation which was acceptable to EPA and execute this protocol once the new stack was erected and the conversion to coal was completed. The company drafted a protocol for the comparative performance evaluation of three models: the NYSDEC model, a modified version of the NYSDEC model (the company's model of choice) and EPA's Complex I model. A series of negotiations then took place between the company, the State of New York and EPA where the details of the protocol and the proposed monitoring network were changed several times. A final protocol[11,12] was agreed upon by all parties in September 1984.* The data base collection phase is not yet underway. It should be completed by 1988.

2.4.2 Preliminary Analysis

The preliminary analysis of the proposed application, contained in the protocol documents, provides a complete description of the existing

---

*Although an appropriate protocol was agreed upon by the source and the control agencies, the construction of the 475-foot stack and conversion of the plant to coal has not yet begun, pending the outcome (final Federal Register approval or disapproval) of the proposed SIP revision.

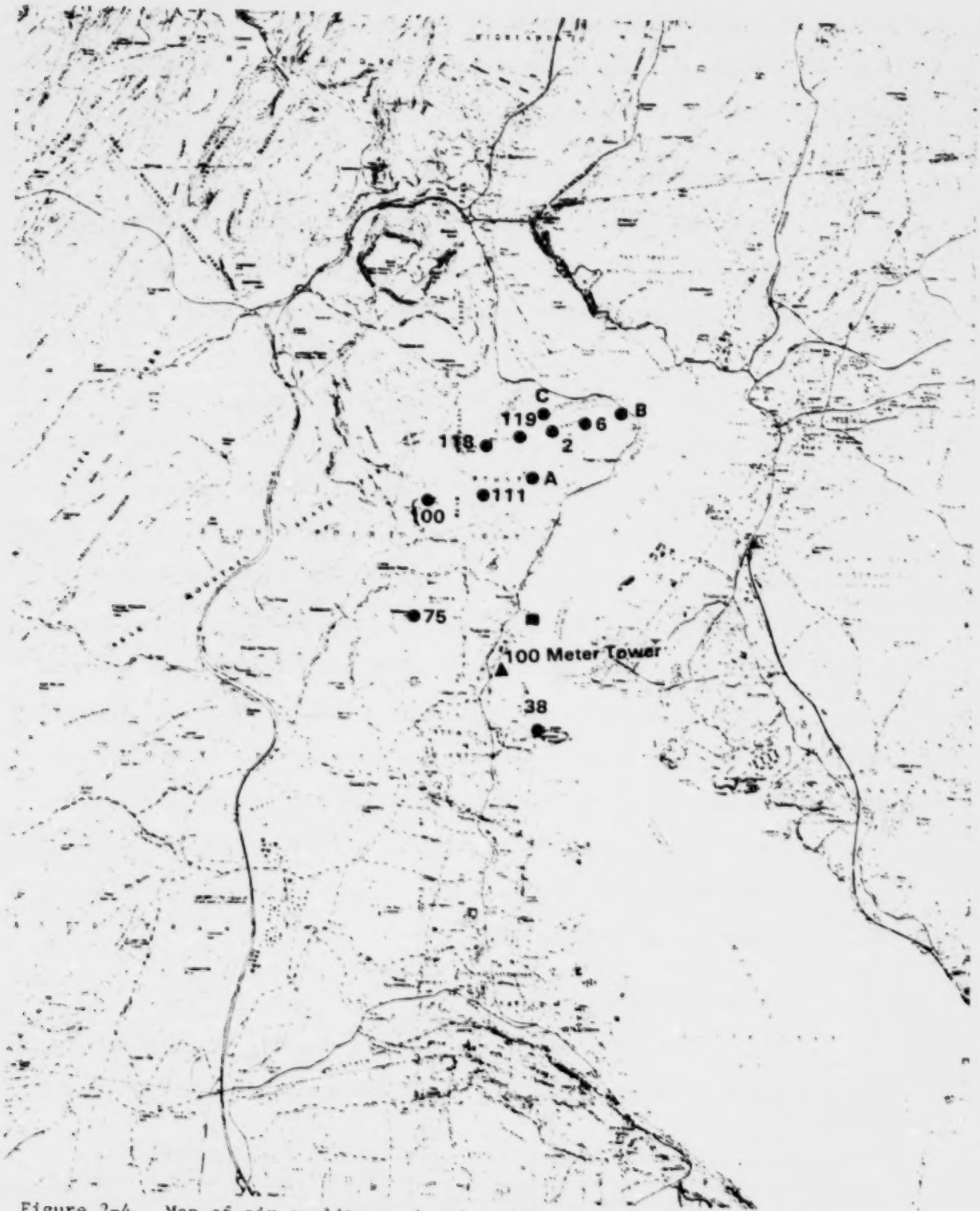


Figure 2-4. Map of air quality monitoring stations (●) and the primary meteorological tower (▲) in the vicinity of the Lovett power plant (■). Ten-meter meteorological towers are also located at Sites 75, 100, 119 and 6.

and proposed source and the surroundings. The regulatory constraints had been established earlier, namely attainment of the SO_2 NAAQS, primarily the short-term NAAQS, on nearby elevated terrain above stack height. The protocol document identifies Complex I as the reference model and the two proposed models, NYSDEC and Modified NYSDEC model. The technical features of the two proposed models are described but no user's manuals are provided. The preliminary analysis does not include a formal technical comparison of the proposed and reference models following the procedures outlined in the Workbook for Comparison of Air Quality Models.

Preliminary estimates of 3- and 24-hour SO_2 concentrations have been made with Complex I and the Modified NYSDEC model, using one year of meteorological data from a tower located at the nearby Bowline power plant. Modeling has been performed for both maximum and average load conditions. The protocol document contains a fairly comprehensive analysis of the results including isopleth maps of maximum short-term concentrations and tables listing the magnitude and locations of the "high-50" estimates. The analysis shows that maximum concentrations for both models would be expected on Dunderberg Mountain to the north of the plant. Complex I estimates are as much as an order of magnitude higher than the Modified NYSDEC model estimates. Secondary maxima are estimated to occur on other more distant terrain features in several directions but these estimates are much lower than those on Dunderberg Mountain.

The protocol document identifies the Bowline power plant, 6 km to the south, as another significant source of SO_2, the plume from which could simultaneously (with Lovett) impact Dunderberg Mountain. The contribution from this plant will be quantified, as a function of meteorological conditions, through

utilization of data from the monitoring network obtained prior to the Lovett plant conversion.

2.4.3 Protocol for the Performance Evaluation

The protocol for the comparative performance evaluation of the three competing models, which is detailed in Appendix D, emphasizes accurate prediction of the peak concentrations and the upper end of the frequency distribution. Twenty (20) percent of the weighting in the protocol involves the calculation of performance statistics that characterize each model's ability to reproduce the measured second-high concentrations at the various monitors. Fifty-eight (58) percent of the weighting is assigned to performance statistics that characterize the models' ability to reproduce the measured concentration in the upper end of the observed frequency distribution, namely the high-25 concentrations. In addition, the protocol assigns twenty-two (22) percent of the weighting to performance measures designed to determine how well the models perform for the entire range (all data) of measured/predicted values, broken out into stable and unstable conditions.

The primary performance evaluation measures are the ratios of observed to predicted concentrations, ratios of the observed to predicted variances and the inverse of these ratios. Seventy-eight (78) percent of the weighting is associated with statistics based on the values of these ratios. These statistics are to be calculated for all combinations of data pairings but most often the unpaired data sets and the data sets paired in space only are used. The analysis of the "all data" data set includes statistics that break out performance by stability category. The other twenty-two (22) percent of the weighting is associated with performance

measures designed to characterize correlation, gross variability and the ability of the models to accurately predict observed concentrations during observed meteorological conditions.

The scoring scheme used for most performance statistics is specified by somewhat complicated formulae and the reader is referred to Appendix D for details. The scheme is similar to that used in the Westvaco protocol. Basically, it involves computing ratios of performance measures between the three competing models and bias ratios or variance ratios for each model. These ratios are then combined in various ways to produce a percentage of maximum possible points for each performance statistic. This result is then multiplied by the maximum possible points for that performance statistic to yield a subscore. Subscores are then summed for each model to yield a total score.

Initially, the model with the highest score is deemed to be most appropriate to apply for regulatory purposes. However, the protocol contains some additional procedures to be employed if the chosen model is shown to underpredict the highest concentrations. The procedure, which is based on the unpaired in time and space comparisons, is as follows:

(1) If the average of the highest ten predicted 3- or 24-hour average concentrations is less than the average of the highest ten observed 3- or 24-hour average concentrations, or

(2) If the highest, second-highest predicted 3- or 24-hour average concentration is less than ninety (90) percent of the highest, second-highest observed 3-or 24-hour average concentration,

then the model predictions will be linearly adjusted to correct this regulatory problem. The adjustment factors will be calculated as the minimum needed to eliminate the two conditions of underprediction listed above.

2.4.4 Data Bases for the Performance Evaluation

The data base for the performance evaluation will consist of a network of monitors/meteorological stations specially designed to cover the area of maximum predicted concentration and to fit the needs of the protocol. This data base will consist of eleven monitors, nine of which are to be in the area north of the plant where preliminary estimates indicate that high concentrations would occur (See Figure 2-4). Monitor #38, located south of the plant, will most often be used to determine background. A 100-meter meteorological tower, instrumented at three levels, will be located at the plant site. Ten-meter meteorological towers are included in the network at sites 6, 119, 100 and 75 but data from the 100-meter tower will be used exclusively. For missing data periods, a hierarchy of default data sources is specified in the protocol. These primarily consist of data from other levels on the 100-meter tower. Wind fluctuation (sigma theta) data from the 10-meter height are used to determine stability inputs to Complex I and the NYSDEC model in accordance with the scheme defined in the Regional Workshops on Air Quality Modeling: A Summary Report. Sigma theta data from the 100-meter level are used as direct input to the Modified NYSDEC model. Morning and afternoon mixing heights will be primarily derived from the Albany National Weather Service data. Hourly emission data and stack gas parameters will be derived from continuous in-stack measurements.

2.5 Guayanilla Basin

The Guayanilla Basin is located on the southern coast of the island of Puerto Rico. The area is characterized by coastal plains with hills rising abruptly from the plains (See Figure 2-5). Historically, several industrial sources of SO_2 have operated in the area but most have shut down. The only currently operating sources, and which are relevant to this analysis,

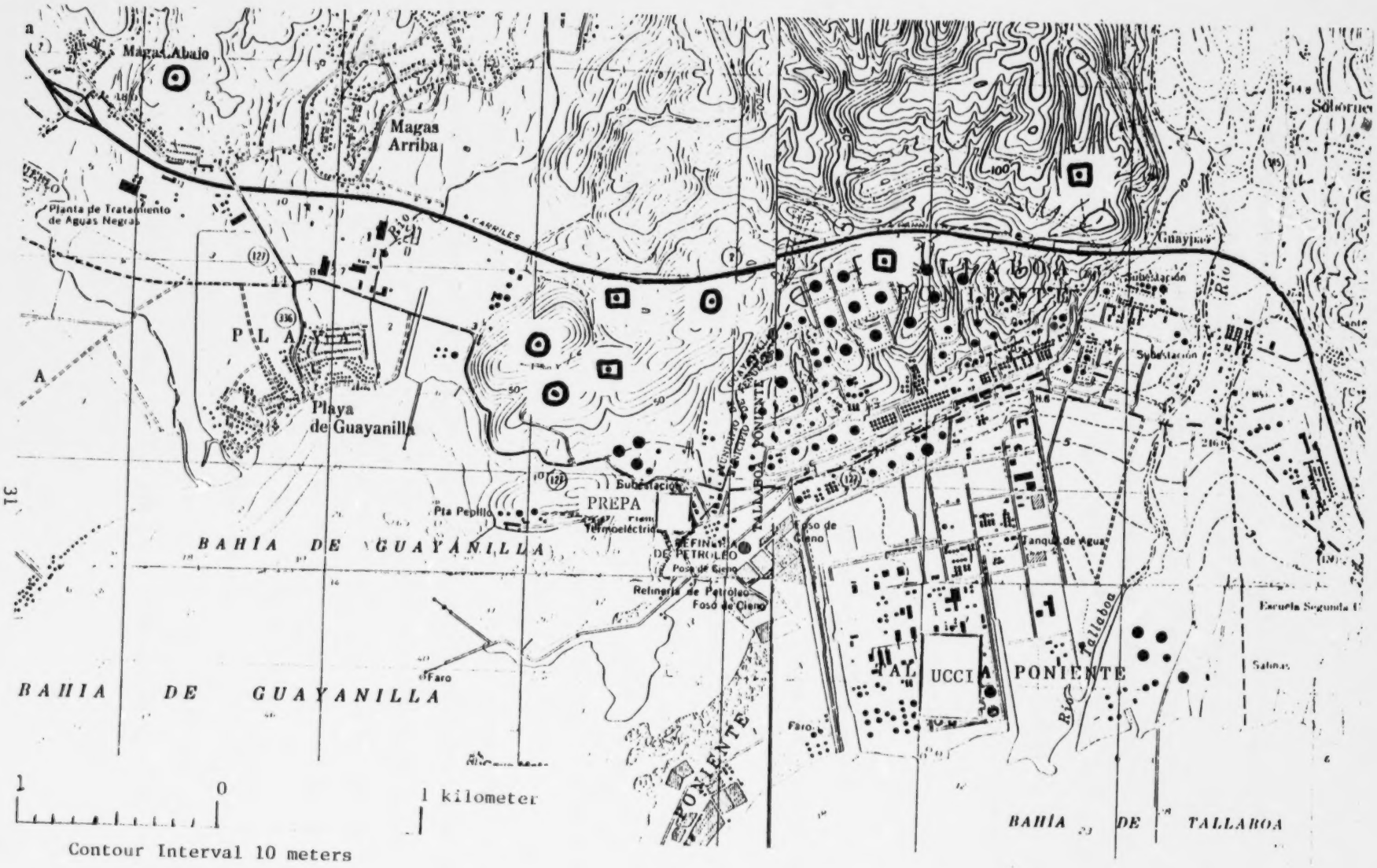


Figure 2-5. Map of existing air quality monitoring network (⊙) and expanded air quality monitoring network (⊡) in the Guayanilla area. The meteorological station is located at the PREPA plant.

are the Puerto Rico Electric Power Authority (PREPA) power plant and the Union Carbide (UCCI) facility, both located near Tallaboa Poniente. The oil fired PREPA plant has stacks ranging in height from 23 feet (7m) to 250 feet (76m) and a combined nominal SO_2 emission rate of 16,545 lb/hour. The UCCI plant has five stacks ranging in height from 38 feet (12m) to 160 feet (49m) with a combined nominal SO_2 emission rate of 1568 lb/hour. Nominal plant grade for both facilities is ten feet (3m) above mean sea level.

2.5.1 Background

The major regulatory concern with these plants has been the attainment of the short-term SO_2 NAAQS on elevated terrain to the north and northwest of the sources. Modeling with EPA's Complex I model indicated that there would be NAAQS violations on the terrain. Industrial interests and the Puerto Rico Environmental Quality Board (PREQB) maintained for several years that emission limits should be based on estimates from the Puerto Rico Air Quality Model (PRAQM), which generally predicts lower concentrations than Complex I. In 1979 Environmental Research and Technology, Inc. (ERT) prepared a report for PREQB entitled "Validation of the Puerto Rico Air Quality Model for the Guayanilla Basin"[14]. The report compared, in various ways, model estimates with historical ambient air quality data from eight monitors (four on elevated terrain) in the area. EPA expressed concerns about the technical aspects of the model and the underestimation of the observed concentrations at some monitors.

In response to these concerns, it was decided in early 1984 that a comparative performance evaluation between the PRAQM and Complex I should be undertaken. Hence EPA developed an analysis and draft protocol for this performance evaluation. The protocol and design of the monitoring

network were then negotiated with PREQB and the industrial interests. A final agreed-upon protocol was issued in December 1984[15].

2.5.2 Preliminary Analysis

The protocol document contains a definition of the regulatory aspects of the problem and a description of the sources and their surroundings. The document states that only the short-term SO_2 concentration estimates are at issue and that the PREQB proposes to use PRAQM in lieu of Complex I to estimate these concentrations, pending the outcome of a comparative performance evaluation. A technical description of PRAQM is contained in the protocol. Apparently no user's manual for PRAQM exists. No formal technical comparison of PRAQM and Complex I, following the procedures outlined in the Workbook for Comparison of Air Quality Models, was performed.

Some preliminary concentration estimates have been made with the PRAQM, Complex I and also the SHORTZ model. The details of these estimates are not provided in the protocol document; however, all parties are privy to the results. The results indicate the following:

1. Maximum concentration estimates occur on elevated terrain to the north and northwest of the plants; however Complex I, PRAQM and SHORTZ all produce different results in terms of magnitude, specific location, and time of the maximum concentrations.

2. Maximum 3-hour and 24-hour concentrations frequently occur both at the monitored locations and in areas that are not monitored.

3. In terms of magnitude, SHORTZ seems to yield the highest concentrations, significantly higher than either Complex I or PRAQM. The PRAQM yields the lowest concentration estimates.

4. The meteorological data indicate a predominance of neutral/unstable conditions associated with the daytime southeast winds. Such conditions generally carry the plumes over the terrain to the northwest of the sources. However, there are occasional hours, during periods of wind shifts, when stable plumes traveling over terrain could have a significant short-term air quality impact.

Based on these results, it has been decided that Complex I is the appropriate reference model and PRAQM the proposed model for the performance evaluation; SHORTZ has been dropped from further consideration. It has also been established that while the existing monitoring network is acceptable for a preliminary performance evaluation, some data from a more detailed network will be necessary to confirm/refute the results of this evaluation. The specifics on how to use the existing network data as well as the design and use of the augmented network and data are discussed below.

2.5.3 Protocol for the Performance Evaluation

The protocol for the comparative performance evaluation of PRAQM and Complex I, which is detailed in Appendix E, specifies that the performance evaluation will be divided into two phases. Phase I is an evaluation for the period January 1983 through December 1984 using monitored data collected at the four existing monitoring sites. Using the selection criteria contained in the protocol a model of choice will be selected in this phase. Phase II of the evaluation is designed to confirm the conclusions reached as a result of the Phase I evaluation. Phase II will be based on six months of air quality data from all eight sites (beginning around September 1984). The specifics of the protocol for each phase are identical, except for a minor stipulation involving the weighting of performance statistics by monitor.

The protocol emphasizes accurate prediction of the peak concentration. Thirty-two (32) percent of the weighting in the protocol involves the calculation of performance statistics that characterize each model's ability to reproduce the measured maximum and second-high concentrations at the various monitors. Sixty-eight (68) percent of the weighting is assigned to performance statistics that characterize the models' ability to reproduce measured concentration in the upper end of the observed frequency distribution, namely the high-25 observed and predicted concentrations.

The primary performance measures are the ratio of the predicted to the observed concentration (average predicted to average observed for the high-25 data set) and the ratio of the variance of predicted concentrations to variance of observed concentrations. Seventy-seven (77) percent of the weighting is on data paired in space but not in time and twenty-three (23) percent on unpaired data. Most performance statistics are to be calculated for 1-, 3-, and 24-hour averaging times. The ratio measures are supplemented by case study statistics, based on the number of cases in common between predicted and observed concentrations (stratified by stability class, for the upper five percent of the 1-hour values).

The Guayanilla protocol specifies that certain performance measures are weighted according to the magnitude of the observed concentrations. The performance statistics for the monitor with a higher observed concentration is given proportionally more weight than that of the next lower ranked monitor. The monitor with the lowest reading receives the least weight.

The scoring scheme used for most performance statistics consists of a percentage of maximum possible points within specified cutoff values. If the performance statistic falls outside of the cutoff values,

no points would be awarded to the model. Within the acceptable range, the percent of possible points is specified in tabular form (discrete values for specified ranges of performance). The tabular values for the bias statistics favor the model that overpredicts, if one model overpredicts to the same extent that the other model underpredicts.

Scores for each model for Phase I and Phase II are determined by totalling the subscores for each performance statistic. For each Phase, the PRAQM is deemed to be the better performer if its score exceeds the score obtained for Complex I by 10 percent. If Phase II leads to a selection of the same model as Phase I, this will be the model for future regulatory use in Guayanilla. If Phase II leads to a selection of a different model, air quality data will be collected for an additional six month period at the eight monitoring sites.

If for both Phases I and II the PRAQM model has a point score at least 10 percent higher than Complex I, it will be considered the preferred model for use in the Guayanilla Basin.

Concentration estimates from the model with the highest score are to be adjusted upward if the highest observed concentrations are significantly underpredicted. The procedure, which is based on the unpaired in time and space comparisons, is as follows:

(1) If the average of the highest ten predicted 3- and 24-hour average concentrations is less than the average of the highest ten observed 3- and 24-hour average concentrations, or

(2) If the highest, second-highest predicted 3- or 24-hour average concentration is less than ninety (90) percent of the highest, second-highest observed 3-or 24-hour average concentration,

then the model predictions will be linearly adjusted to correct for this regulatory problem. The adjustment factors will be calculated as the minimum needed to eliminate the two conditions of underprediction listed above. If Phase II of the evaluation confirms the selection of the model determined by Phase I, but there is a difference in terms of whether an adjustment is warranted or different adjustments are indicated, the adjustment that is most conservative (leads to the most stringent emission limit) will be selected.

2.5.4 Data Bases for the Performance Evaluation

The data base for the Phase I performance evaluation consists of two years of data from an existing 4-station monitoring network and an on-site meteorological tower. The data base for Phase II consists of six months of data from an 8-station network including the original four monitors plus four additional monitors situated to better cover the area of predicted maximum concentration and to fit the requirements of the protocol. The locations of the monitors are indicated in Figure 2-5. Data from the same meteorological tower are used in Phase II.

Sensors are mounted on a meteorological tower, located near the PREPA plant, to collect wind speed, wind direction and temperature data at 10 and 76 meters. Wind data from 76 meters will be scaled to plume height with the 10-meter data used as backup. Wind fluctuation (sigma theta) data collected at 10 meters will be used to determine Pasquill-Gifford stability class for both models according to the scheme described in the Regional Workshop on Air Quality Air Quality Modeling: A Summary Report. Periods of missing data will be eliminated from the performance evaluation. Climatological average daily maximum and minimum mixing heights will be

used. Hourly emission data and stack gas parameters are to be generated from load levels, fuel consumption rates, fuel sampling and other surrogate parameters that are technically defensible.

At the present time data collection from Phase II is still underway and no results from either the Phase I or Phase II performance evaluation are available.

2.6 Other Protocols

In addition to the five major performance evaluation analyses and protocols discussed above, EPA is aware of three other analyses/protocols written to assess the acceptability of proposed models for specific sources. For one reason or another these efforts never reached fruition, i.e. no decisions were made or are intended to be made, on emission limits based on the chosen model. Brief descriptions of these three efforts are provided below.

2.6.1 Example Problem

One such effort is the example problem which illustrates the use of the Interim Procedures for Evaluating Air Quality Models (Revised) and is included as Appendix B to that document. This narrative example was based on 1976 emissions data from the Clifty Creek power plant in Indiana and 1976 SO_2 ambient data from a 7-station network in the vicinity of the plant.

The narrative example was specifically designed to illustrate in a very general way the components of the decision making process and the protocol for performance evaluation. As such, the preliminary technical/regulatory analysis of the intended model application, while included in the example, was significantly fore-shortened from that which would normally be needed for an actual case. Also, since the evaluation was carried out

on an existing data base, the example did not illustrate the design of the field measurement program required to obtain model evaluation data.

The example problem protocol incorporated a broad spectrum of performance statistics with associated weights. The number of statistics contained in the example was overly broad for most performance evaluations and perhaps, even for the problem illustrated. Thus its use was not intended to be a "model" for actual situations. For an individual performance evaluation it was recommended that a subset of statistics be used, tailored to the performance evaluation objectives of the problem. Similarly, the method used to assign scores to each performance statistic (non-overlapping confidence intervals) was not intended to be a rigid "model" but only an illustration of one of several possible techniques to accomplish the goal.

2.6.2 Gibson Power Plant

In May 1981, Public Service Company of Indiana (PSI) submitted to the Indiana Air Pollution Control Division (IAPCD) a report[16] which outlined proposed procedures for conducting a comparative performance evaluation of models applicable to setting the SO_2 emissions limit for the Gibson power plant. PSI proposed to establish a monitoring network (actually augment an existing network), the data from which would be used to establish whether either of two versions of the MPSDM model would be more appropriate to apply to the plant than EPA's CRSTER model. The report contained an incomplete performance evaluation protocol that would be used in the evaluation.

Following submittal of this report, a series of negotiations on the protocol and the monitoring network took place between PSI and IAPCD. Some of these negotiations involved EPA. In July 1981, IAPCD accepted the PSI plan, but EPA continued to express major concerns about the technical aspects of the proposed models, on the monitoring network and on the

protocol. These concerns were not resolved and in June and August 1982 EPA sent letters to PSI[17,18] cautioning them that the Agency could not accept the results of the performance evaluation, if the company chose to proceed.

Apparently PSI proceeded with the evaluation and collected the one year of data from the network. The outcome of the evaluation is unknown at the present time.

2.6.3 Homer City Area

In November 1982, the Pennsylvania Electric Company (Penelec) submitted to the State of Pennsylvania Department of Environmental Resources (DER) a report, "Protocol for the Comparative Performance Evaluation of the LAPPES and Complex I Dispersion Models Using the Penelec Data Set"[19]. The company's intent was to execute the protocol and demonstrate the acceptability of the LAPPES model in the Homer City, Pennsylvania area so that this model could be used to revise SO_2 regulations for four area power plants. The plants, which have varying stack heights, are located in moderately complex terrain with receptors of concern located both above and below the heights of the stacks.

The protocol was reviewed by DER and by EPA and a number of comments/suggestions were provided to Penelec. The most significant comment involved the choice of Complex I as the only reference model. An examination of the topography in relationship to stack heights in the area revealed that many of the monitors (and most of the terrain) were below most of the physical stack heights. In fact, when expected plume rise was considered, only the Seward plant, because of its relatively short stack, exhibited a real risk of direct, stable plume impaction on terrain; the Conemaugh plant was somewhat marginal in this regard. From an overall performance evaluation standpoint, this resulted in a dilemma. Some of the monitors were considered

"flat terrain" receptors for which CRSTER was the appropriate reference model while some were complex terrain sites where Complex I might be appropriate. An added complexity was that, because of varying stack heights, some monitors might be both flat terrain and complex terrain receptors depending on which power plant was being modeled. Thus, Complex I was not the appropriate model for all monitors, as proposed in the protocol and it will likely underestimate concentrations at receptors that are below stack height.

Although the protocol has been executed,[20] the issue regarding the choice of an appropriate reference model(s) has apparently never been resolved.

3.0 INTERCOMPARISON OF APPLICATIONS

In this section, the details of the five major applications of the Interim Procedures are intercompared. Each subsection below corresponds roughly to and in the same order as Sections 2, 3 and 4 (and inherent subsections) of the Interim Procedures for Evaluating Air Quality Models (Revised). It is also possible to identify the subsections below with sequential blocks in the flow chart for the Interim Procedures provided in Figure 1-1 above. In this way it is possible to analyze the five applications according to subject matter as it appears in the Interim Procedures.

In the subsections below the common and differing features among the five major applications are described. Where appropriate, these features are compared to recommendations/suggestions contained in the corresponding section/subsection of the Interim Procedures and similarities/differences are noted.

The material contained in this section is intended to be factual, i.e. additional interpretation or opinion is generally avoided. Any interpretations and/or opinions that are provided are only intended to reflect the views, or apparent views, contained in the individual protocols and related documents.

3.1 Preliminary Analysis

The Interim Procedures recommend that before any performance evaluation protocol is written or any performance data are identified/collected, the applicant should conduct a thorough preliminary analysis of the situation. This analysis serves to describe the source and its environment, the regulatory constraints, the proposed and reference

models, the relative technical superiority of each model's attributes, and the ambient consequences of applying each of the models to the regulatory problems.

In each of the five applications some preliminary analyses were conducted, although the level of detail and the degree of adherence to the recommended procedures varied considerably as described in Sections 3.1.1 - 3.1.4 below.

3.1.1 Regulatory Aspects

The Interim Procedures recommend that each application identify the regulatory aspects of the problem including the pollutant, averaging times and applicable regulations.

In each of the five applications this aspect of the analysis was quite thoroughly covered. All of the applications involved SO_2 emitters and the NAAQS were identified as the limiting regulatory constraint, i.e. PSD increments or other ambient criteria did not apply. In most applications compliance with the 3- and 24-hour SO_2 NAAQS was identified as the relevant regulatory requirement although it was tacitly implied that the model of choice ultimately would also be used to demonstrate compliance with the annual standard. The one exception was the Baldwin power plant where it was established that the proposed model, if determined to be acceptable, would only be used to estimate 3-hour concentrations.

3.1.2 Source Characteristics and Source Environment

The Interim Procedures recommend that each proposed evaluation be accompanied by a complete description of the source(s) and its surroundings.

Table 3-1 compares the sources/environment for the five major protocols. The Table shows that power plants were the major sources of SO_2

Table 3.1 Source Characteristics and Source Environment

	BALDWIN	WESTVACO	WARREN	LOVETT	GUAYANILLA
TYPE OF SOURCE	POWER PLANT	PULP MILL	POWER PLANT	POWER PLANT	POWER PLANT INDUS. PLANT
NO. OF STACKS	3	1	1	1	12
STACK HEIGHT (FT)	605	623	200	475	13-250
TOTAL EMISSIONS (LB/HR)	101,588	4083	2420	3837	18,203
NEARBY TERRAIN HEIGHT[1] (FT)	INSIGNIFICANT	1500	700	1100	500
URBAN/RURAL	RURAL	RURAL	RURAL	RURAL	RURAL
BACKGROUND	MONITORED	MONITORED	MODELED & MONITORED	MONITORED	ASSUMED TO BE ZERO

1. Nominal height above plant grade.

emissions with the exception of Westvaco. Most evaluations involve, effectively, a single tall stack. The Guayanilla evaluation is the only evaluation involving a true multiple stack situation. In all cases except Baldwin, complex terrain is a major consideration with nearby terrain well above stack height(s). All of the sources are in a rural environment and most are isolated from any neighbors, i.e. the contribution from nearby sources is not considered to be significant. The exceptions are Warren, where a nearby plant is to be explicitly modeled, and Lovett, where the contribution from the nearby Bowline power plant will be determined from monitoring data.

3.1.3 Proposed and Reference Models

The Interim Procedures state that for each evaluation it is highly desirable to choose a proposed and a reference model applicable to the situation. (For cases where no reference model can be identified, the Interim Procedures suggest an alternative approach that can be used to determine acceptability of the proposed model.) It is further recommended that each model be well documented, by a user's manual if possible. The technical features of the competing models should be intercompared, preferably using techniques described in the Workbook for Comparison of Air Quality Models.

Table 3-2 lists the proposed and reference models for each of the five evaluations and the degree to which these models are documented and intercompared. The Table shows that each evaluation involves a different proposed model. In the case of Lovett there are two proposed models. The Complex I model is most often used as the reference model. All of the preliminary analyses contain technical descriptions of the models to be evaluated as well as technical/descriptive comparisions of the relavant

Table 3.2 Proposed and Reference Models

	BALDWIN	WESTVACO	WARREN	LOVETT	GUAYANILLA
PROPOSED MODEL	MPSDM	LUMM	LAPPES	NYSDEC, MOD. NYSDEC	PRAQM
TECHNICAL DESCRIPTION	YES	YES	YES	YES	YES
USER MANUAL	YES	NO[1]	NO[1]	NO	NO
TECHNICAL COMPARISON WITH REFERENCE MODEL	YES	YES	YES	YES	YES
REFERENCE MODEL	CRSTER	SHORTZ	COMPLEX I	COMPLEX I	COMPLEX I
TECHNICAL DESCRIPTION	YES	YES	YES	YES	YES
USER MANUAL	YES	NO[1]	NO[2]	NO[2]	NO[2]

1. A User Manual for the model exists but the model was modified for use in the application

2. Complex I is self-documented computer code for which no specific user manual exists. However the model is a modification of MPTER, for which a user manual does exist.

features of the models. In only one case was the "Workbook" comparison rigorously applied. Explicit, up to date, user's manuals were most often not available. In some cases such manuals did exist but were not up-to-date with the version of the models to be used in the evaluation.

3.1.4 Preliminary Concentration Estimates

The Interim Procedures suggest that preliminary concentration estimates be obtained from both the proposed and the reference models, as an aid to writing the protocol and designing the requisite data bases.

In the three most recent protocols (Warren, Lovett and Guayanilla) such estimates were made, although they are not well documented for Guayanilla. In the Baldwin and Westvaco evaluations, it is not evident that any formal estimates were made although both the source and the control agencies had a good idea of the consequences (location and magnitude of high estimates) of applying the models.

3.2 Protocol for the Performance Evaluation

The Interim Procedures require that a protocol be prepared for comparing the performance of the reference model and proposed model. The protocol must be agreed upon by the applicant and the appropriate regulatory agencies prior to collection of the requisite data bases.

In each of the five cases such a protocol was written, negotiated with the control agencies, and a final protocol to be used in the evaluation was established. The relative details of the various protocols are compared in the following subsections, 3.2.1 - 3.2.5. The degree to which the negotiating parties were in full agreement that the final established protocol was optimum is discussed in Section 3.4.

3.2.1 Performance Evaluation Objectives

The Interim Procedures suggest that the first step to developing a model performance protocol is to translate the regulatory purposes associated with the intended model application into performance evaluation objectives which, in turn can be linked to specific data sets and performance measures. Ranked-order performance objectives are suggested with the primary objective focussing on what is perceived to be (from the preliminary analysis) the critical source-receptor relationship, i.e. the averaging time, the receptor locations, the set(s) of meteorological conditions and the source configuration that are most likely associated with the design concentration. Lower-order objectives, e.g. second-order, third-order, etc., would focus on other source-receptor relationships which must be addressed when the chosen model is ultimately applied to the situation, but are not perceived to be of prime importance (not as likely to be associated with a design concentration) when the chosen model is applied.

In the five protocols, specific sets of ranked-order objectives were not stated, at least in the sense described above. However, it is apparent from the choices of data sets and performance measures, the weighting of data sets/performance measures, the sometimes-used differential weighting of individual monitor data, and the scoring schemes employed in the protocols, that the writers had such ranked-order objectives implicitly in mind. Most of the protocols explicitly stated a single broad objective which focuses on an accurate prediction of peak short-term concentrations. These statements were generally not narrowed down to include specific receptor locations, the importance of time pairing or critical meteorological conditions. However, as mentioned above, it is evident from the protocols' contents that these single broad performance objective statements really did implicitly contain sets of ranked-order specific objectives.

3.2.2 Data Sets, Averaging Times and Pairing

The Interim Procedures mention a number of possible data sets which can be considered but makes no specific recommendation as to the choice of data sets for an individual situation.

Table 3-3 compares the data sets contained in each of the five major protocols and the weighting (percent of maximum possible points) of each data set. The protocols are arranged roughly chronologically across the top of the table, in the order in time when each was finalized, to see if there are any trends in the choices of data sets or weighting. No obvious pattern is apparent. It is clear from Table 3-3 that each of the protocols focuses on the common broad performance evaluation objective of accurate prediction of peak short-term concentration. However, it is obvious from the choice and weighting of data sets that the protocol writers had different ideas on how to best meet that objective. Three of the five protocols examined the highest observed/predicted concentration data set as well as the second-highest data set. All of the protocols tested the competing models against the second-highs and the high-25 set, although there were considerable differences in the weighting among the protocols. Two protocols specify that some performance statistics will be calculated for all data but the weighting of this data set is lower than the peak/high-25 data sets.

The Interim Procedures suggest that performance of models whose basic averaging time is shorter than the regulatory averaging time should be evaluated for that shorter period as well as averaging times corresponding to the regulatory constraints.* Since all five cases involved SO_2 models

*Most models compute sequential concentrations at each receptor over a short time average, e.g., 1-hour. Average concentrations for longer periods, e.g., 3-hour, 24-hour, are arrived at by summing the sequential short-term averages.

Table 3-3 Weighting (%) of Maximum Possible Points by Data Set, Averaging Time and Degree of Pairing

	BALDWIN	WESTVACO	WARREN	LOVETT	GUAYANILLA
DATA SET					
MAXIMUM	0	21	15	0	12
SECOND HIGH	55	22	28	20	15
HIGH 25	45	57	43	58	74
ALL DATA	0	0	14	22	0
AVERAGING TIME					
1-HOUR	0	19	20	36	34
3-HOUR	100	37	30	36	27
24-HOUR	0	37	36	26	39
ANNUAL	0	7	14	2	0
PAIRING					
UNPAIRED	70	32	60	44	23
SPACE ONLY	5	62	40	34	77
TIME ONLY	0	3	0	21	0
SPACE AND TIME	25	3	0	1	0

whose basic averaging time is one hour, this would suggest that 1-hour statistics be calculated as well as 3-hour, 24-hour, etc. Table 3-3 shows that all of the protocols except Baldwin specify that performance statistics should be calculated for 1-, 3- and 24-hour averaging times. For Baldwin it was established up-front that the proposed model, if selected, would only be applied for the 3-hour averaging time. This may be the reason why statistics are not to be calculated for other averaging times, including 1-hour. Computation of the annual concentration is not a significant issue in any of the cases. This is apparently the reason that low or no weight is given to statistics for that averaging time.

Weighting may also be distributed according to performance statistics calculated for data paired in space, time, both space and time or completely unpaired. The Interim Procedures discuss the various possible degrees of pairing asociated with each data set but makes no specific recommendation as to which to choose or the weighting distribution. Instead, the Interim Procedures suggest that through the development of performance evaluation objectives, pairing can be identified.

Table 3-3 also shows the weighting of maximum possible points according to the degree of pairing specified in each of the five protocols. Since detailed performance evaluation objectives are generally lacking for these protocols, it is difficult to establish a rationale for the seemingly significant variation of weighting among the protocols. In each evaluation a relatively isolated point source of SO_2 controlled the short-term ambient SO_2 levels in its vicinity. Thus it is not very important that the models predict the concentration in time and space accurately; only the magnitude is of importance. This suggests that completely unpaired performance statistics would be of prime importance. Table 3-3 shows that unpaired statistics

were important in all five protocols but the weighting and degree of importance vary significantly. In fact, in the Westvaco and Guayanilla protocols, data paired in space only seem to be regarded as the most pertinent. Although specific rationales are generally lacking, it appears that the protocol writers were concerned with model credibility. Credibility in model performance can be linked to the ability of the models to reproduce measured concentrations at the right place, right time and perhaps both. This explains (perhaps) the varying degree of pairing.

3.2.3 Performance Measures

The Interim Procedures state that the basic tools used in determining how well a model performs in a given situation are performance measures. These performance measures are viewed as surrogate quantities whose values/statistics serve to characterize the discrepancy between predictions and observations.

Table 3-4 lists, by data set, the various performance measures used in the protocols for characterizing performance for that data set. From an overall perspective the Table seems to indicate that, while there are some similarities, there are also a wide variety and combinations of performance measures used among the protocols. Each protocol seems to contain a more or less unique combination of measures used to characterize performance and this combination often differs from those suggested in the Interim Procedures. Some of the protocols contain certain performance measures not mentioned in the Interim Procedures. For example, three of the protocols contain a performance measure, M_c*, designed to test the models'

---

*M_c is not a unique performance measure but refers to schemes for quantifying this type of performance which differ among the various protocols. See Appendices B, D and E for details.

Table 3-4 Performance Measures Used in the Protocols

DATA SET	BALDWIN	WESTVACO	WARREN	LOVETT	GUAYANILLA
PEAK VALUES (HIGH & SECOND HIGH)	d	d	$\bar{d}$ $\lvert\bar{d}\rvert/\bar{C}_o$ C_p/C_o R	C_p/C_o, C_o/C_p	C_p/C_o
HIGH-25	d S_d $RMSE_d$ M_c	$\lvert\bar{d}\rvert$ $s_p^2/s_o^2, s_o^2/s_p^2$	C_p/C_o s_p^2/s_o^2 By Stability: C_p/C_o s_p^2/s_o^2	$\bar{C}_p/\bar{C}_o$, $\bar{C}_o/\bar{C}_p$ s_p^2/s_o^2, s_p^2/s_p^2 M_c	$\bar{C}_p/\bar{C}_o$ s_p^2/s_o^2 M_c
ALL DATA	F		C_p/C_o $\bar{d}$ $\lvert d\rvert/C_o$ R	$\bar{C}_p/\bar{C}_o$, $\bar{C}_o/\bar{C}_p$ By Stability: C_p/C_o, C_o/C_p s_p^2/s_o^2, s_o^2/s_p^2 R d^2	

d = Residual

S_d = Standard deviation of residual

S_p^2 = Variance of predicted concentration

S_o^2 = Variance of observed concentration

F = Frequency distribution

R = Correlation coefficient

M_c = Meteorological cases in common

$RMSE_d$ = Root-mean-square-error of the residual

capability to reproduce observed concentrations during observed meteorological conditions. Note also that certain performance measures such as the correlation coefficient, the variance of the residuals and statistics on the frequency distribution were not widely used in the protocols. Where they were used, they were not weighted heavily.

One specific point revealed by Table 3-4 concerns the use of performance measures that characterize the model bias. The Interim Procedures suggest that model bias is an important quantity in performance evaluations and that the model residual is an appropriate measure to characterize the bias. In the earlier protocols, Baldwin and Westvaco, the model residual was used exclusively in this regard. However, in time, as indicated by the more recent protocols on the right side of Table 3-4, the residual is used less frequently or not at all. Instead, various combinations of the ratio of the predicted to observed concentration are used to characterize the bias. No clues/rationale are contained in these recent protocols that suggest a reason for using ratios instead of residuals. At the time that these protocols were negotiated with the control agency, no significant objections were apparently raised over the use of ratios in lieu of residuals.

3.2.4 Model Performance Scoring

One of the more difficult aspects of writing a performance evaluation protocol is devising a scheme which, for each performance measure or other surrogate measure, objectively quantifies the degree to which the model reproduces measured concentrations. The specification of the details of this concept, which is called "scoring" in the Interim Procedures, lacks a clear technical basis or a basis in past experience. The Interim Procedures

recognize this lack of guidance and invite the use of innovative schemes, although the use of confidence intervals is mentioned as one such possible scheme.

The lack of guidance in this area is well reflected by the wide variety of scoring schemes that are specified in the various protocols. In fact, each protocol generally contains several different schemes in itself. No attempt is made here to intercompare the details of the various scoring schemes employed; the reader is referred to Appendices A-E and the specific protocols in the Section 5.0 References for these details.

In general, most of the schemes ultimately generate what might be termed a performance factor. The performance factor is either a measure or an indicator of how well the model performs in relation to measured data. The method (usually formulae) used to arrive at the performance factor depends on the specific measure of performance (residual, ratio, correlation coefficient, etc.) and varies widely among the protocols. The performance factor, once obtained, is either multiplied by the maximum possible points to obtain a subscore for that performance measure, or a table is entered which provides point subscores for specific ranges of the performance factor. The tables most often have "cutoff values" above or below which a zero subscore is specified. For measures of the bias, the table is sometimes skewed in favor of overprediction, i.e. a given amount of overprediction is awarded more points than the same degree of underprediction.

Once the various subscores are obtained, they are, in each protocol, summed to obtain a total score for each model. In some cases, the performance factor mentioned above involves performance statistics from both models. Thus in these cases the scores obtained for each model are

not truly independent indicators of how well the model performs relative to measured data but contain some elements of relative performance between models. In any event, the scores for each model are then compared to obtain a preliminary indication of which model is the better performer.

At this point the Interim Procedures suggest that, among other things, it might be desirable to define a "window" of marginal performance. If the apparently better performer falls in the window then the results of the technical evaluation could be used to arrive at a final decision. In only one of the protocols, Guayanilla, is the window concept used and in that case it is merely stated that the proposed model, if it receives a higher score, will not be chosen unless that score exceeds that of the reference model by ten percent.

The Interim Procedures also suggest that it might be undesirable to apply the chosen model should it be shown to underpredict critical high concentrations. In this case it is suggested that the chosen model be "corrected" or "adjusted" to the degree which it apparently underpredicts. In the three most recent protocols (Warren, Lovett, and Guayanilla) this concept is employed, although the details on the criteria for and the method of correcting the model estimates vary.

3.3 Data Bases for Performance Evaluations

The Interim Procedures suggest that three types of data bases can be used for performance evaluation purposes, data from an on-site specially designed network, data from an on-site tracer experiment and, rarely, data from an off-site network. The five performance evaluations utilize data from an on-site network of SO_2 monitors and other instruments. Table 3-5 shows that three of those networks were specially designed for the performance

Table 3-5 Data Bases For Performance Evaluations

DATA	BALDWIN	WESTVACO	WARREN	LOVETT	GUAYANILLA
TYPE OF NETWORK	SPECIAL DESIGN	EXISTING	SPECIAL DESIGN	SPECIAL DESIGN	PHASE I- EXISTING PHASE II- SPECIAL DESIGN
NO. OF MONITORS	10	9	7	11	PHASE I-4 PHASE II-8
LENGTH OF DATA RECORD	1 YEAR	1 YEAR	1 YEAR	1 YEAR	PHASE I-2 YEARS PHASE II- 6 MONTHS
NO. OF ON-SITE MET. TOWERS*	1	2	2	5	1
ON-SITE MET. DATA	WD,WS,WF,T	WD,WS,WF,T	WD,WS,WF,T	WD,WS,WF,T	WD,WS,WF,T
OFF-SITE MET. DATA	NWS: MXHT & STABILITY	NONE	NWS: MXHT & MISSING DATA PERIODS	NWS: MXHT	CLIMATOLOGICAL MXHT
EMISSIONS DATA	LOAD LEVEL/ FUEL SAMPLES	IN-STACK DATA	LOAD LEVEL/ FUEL SAMPLES	IN-STACK DATA	LOAD LEVEL/ FUEL SAMPLES

WD = Wind Direction
WS = Wind Speed
WF = Wind Fluctuation or Turbulence Intensity
T = Temperature
NWS = National Weather Service
MXHT = Mixing Height

* Data from only one primary tower are used, except for data substitutions when primary data source is not operating

evaluation. The Westvaco protocol utilizes data from a network that was originally designed to monitor compliance of the source with the NAAQS. As pointed out in Section 2.2, this network was judged to be acceptable for performance evaluation purposes. In the Guayanilla protocol, the existing network of four monitors was judged to be only marginally adequate for performance evaluation purposes. Thus there will be a Phase II performance evaluation, where six months of data from an augmented, specially designed network are to be utilized.

The Interim Procedures recommend that the number and spatial coverage of monitors is a tradeoff between the scientific desire for wide coverage with a dense array and the practical constraints of cost and logistics. In any event the requisite network must have sufficient spatial coverage/density to address important source-receptor relationships identified in the preliminary analysis and to meet the needs of the protocol. Table 3-5 shows that the networks to be utilized for performance evaluations contain about the same number of ambient monitors, i.e. ranging from 7 to 11. Further investigation of these networks reveals that in each of them nearly all of the monitors are fairly densely clustered in the area of expected maximum concentration with one or two monitors, generally to be used for assessing background, located well outside of this area.

The Interim Procedures suggest that a 1-year data collection period is normally the minimum in order to calculate performance statistics that are related to the NAAQS, i.e. the high second-high concentration. Table 3-5 shows lengths of record ranging from one to 2-1/2 years will be used for performance evaluation purposes.

In all of the performance evaluations the primary source of meteorological data is from an on-site network. Although some of the networks con-

tain multiple towers (See Table 3-5), none of the models to be considered in the evaluations is capable of utilizing spatially divergent meteorological data inputs. Thus, meteorological data inputs to the models are pre-specified to be from a single tower, with other stations used as backup for missing data periods. In most cases, on-site wind fluctuation data (sigma-theta) are to be used either as direct input to the models or as a means to categorize stability. Mixing heights are usually derived from off-site National Weather Service temperature sounding data. On-site temperature data are sometimes used to interpolate hourly values of the mixing height.

The Interim Procedures recommend that in-stack instrumentation is the preferable data source to be used in deriving hourly emissions and values of stack gas parameters. Table 3-5 shows that such in-stack instrumentation is or will be in place at Westvaco and Lovett. The other three performance evaluations derive emissions and stack data from surrogate measures such as fuel analyses and documented load level information.

3.4 Negotiation of the Procedures to be Followed

The Interim Procedures strongly recommend that the applicant (source) maintain close liaison with the reviewing agency at the beginning and throughout the project. It is especially important that the protocol and design of the monitoring network be negotiated and agreed upon before any data are in-hand. In each of the five cases, such negotiations took place. These negotiations generally took place at two points of time: (1) in advance of any work on the project itself where the need to do a comparative model evaluation was identified as an acceptable way to resolve differences of opinion on model acceptability and (2) after a draft protocol was written and the proposed network was designed or identified.

Table 3-6 lists the major components of the model evaluation process as identified in the Interim Procedures and as discussed in Section 3.1-3.3 above. For each of the five cases the Table indicates whether that component was a significant or minor issue in the negotiation process. A significant issue is defined as one where there was a significant difference of opinion between the source and the control agency or, in some cases, between control agencies. A minor issue is one where the source did not strongly object to the control agency's request for changes or additions to the analyses, protocol or data base collection plans. (A minor issue may have resulted in a significant amount of additional analysis). If no entry is made in the Table it indicates that there was no issue or that the component was apparently not discussed in the review process.

The Table shows that regulatory aspects and the design of the data base network were significant issues common to all of the projects. The resolution of these issues was, in fact, the decision to go ahead with the model evaluation. The network design issues generally reflect Agency concerns that monitors be located in areas of expected maximum concentration. It is interesting to note that, in spite of the wide variation in the details of the protocol, discussed in Section 2.2, there was apparently not much debate on these details.

Table 3-6 Issues Involved in Negotiations

	BALDWIN	WESTVACO	WARREN	LOVETT	GUAYANILLA
PRELIMINARY ANALYSIS					
REGULATORY ASPECTS	S	S	S	S	S
SOURCE & SOURCE ENVIRONMENT	-	-	-	-	-
CHOICE OF PROPOSED MODEL	-	-	-	M	M
DOCUMENTATION OF PROPOSED MODEL	-	-	M	-	-
CHOICE OF REFERENCE MODEL	-	S	-	-	S
PRELIMINARY ESTIMATES	-	-	M	M	M
PROTOCOL					
PERFORMANCE EVALUATION OBJECTIVES	-	-	-	-	-
CHOICE OF DATA SETS	S	-	-	-	-
CHOICE OF AVERAGING TIME	S	-	-	-	-
DEGREE OF PAIRING	-	-	-	-	-
CHOICE OF PERFORMANCE MEASURES	S	-	-	-	-
WEIGHTING (DISTRIBUTION)	S	-	M	-	-
WEIGHTING OF MONITORS	-	S	-	-	-
SCORING	-	-	-	-	-
ADDITIONAL CRITERIA[1]	-	-	M	M	M
DATA BASES					
NETWORK DESIGN	S	S	S	S	S
NUMBER OF MONITORS	-	-	-	-	S
CHOICE OF METEOROLOGICAL INPUTS	-	S	M	M	S

M = Minor difference of opinion
S = Significant difference of opinion
- = No difference of opinion stated

Footnote

1. Includes criteria to guard against underprediction of critical concentrations.

4.0 FINDINGS AND RECOMMENDATIONS

The summaries and analyses of several major cases, which utilize guidance contained in the Interim Procedures for Evaluating Air Quality Models, lead to the following general findings. These findings parallel the basic principles of the Interim Procedures listed in Section 1.2.

Finding 1. Up-front negotiations between the applicant and the regulatory agencies on the nature of the protocol and design/utilization of the data base network took place in each case. Up-front discussions on the preliminary analysis did not always take place. This lack of early communication sometimes led to backtracking to fill in needed analyses.

Recommendation

In the interest of expediency, the applicant should initiate frequent discussions with all of the control agencies that will be ultimately involved in reviewing/approving the evaluation. Based on experience it is especially important that discussion take place before the preliminary analysis is conducted such that the applicant can provide all the relevant information required for the case.

Finding 2. For each case a detailed protocol for performance evaluation was written.

Recommendation

Establishing an up-front protocol has worked very well as the central mechanism for decision-making on the appropriate model. This should be continued.

Finding 3. For each case an on-site data base network was established or identified as meeting the needs of the protocol and the technical/regulatory requirements. In three of the evaluations a network was specially designed to meet these needs. In one evaluation the existing network was augmented

to meet these needs. In one evaluation the existing network was judged to be adequate without any modification.

Recommendation

It is clear from experience that it is highly important to establish the design of the data base network before any data are collected or at least before any data are available to the user. This practice should be continued.

Finding 4. Details of the protocol and network design were well documented in each case. However, details of the preliminary analysis and the negotiation process were not always well documented.

Recommendation

It has become increasingly obvious that the preliminary analysis, especially the preliminary concentration estimates, plays an important role in the design of the protocol and the data base network. In the interest of avoiding misunderstanding, complete documentation of this preliminary analysis is strongly recommended.

Finding 5. For the two cases where the evaluations have been completed, the decision on the appropriate model was made as prescribed in the protocol.

Recommendation

The execution of an established protocol leading to a rationalized decision on the more appropriate regulatory model is a basic premise in the Interim Procedures. This practice should be continued.

Other more specific findings and recommendations are:

Finding 6. Each of the five protocols involved large point sources of SO_2 where attainment of the short-term NAAQS was at issue. Four of the sources were located in complex terrain.

Recommendation

The use of the Interim Procedures over a broader range of model evaluation problems is encouraged.

Finding 7. In each of the five cases a technical description of the proposed model was provided. However, a rigorous technical comparison of the proposed and reference models, according to procedures outlined in the Workbook for Comparison of Air Quality Models, was not generally performed. Also, user's manuals on both the proposed and reference models were generally not available.

Recommendation

The results of rigorous application of the "workbook" procedures have not been used as decision criteria for any of the cases covered in this report. However, it is important that the technical features of the competing models be compared and the workbook provides a good framework for making such comparisions. Thus its continued use, at least in the latter regard, is encouraged.

Either a self-documenting code or a user's manual should be provided for each model under consideration. All data bases used in the evaluation should be provided.

Finding 8. Preliminary estimates of expected concentration levels were made in some cases; these results were not always well documented.

Recommendation

Preliminary estimates should be submitted in all future applications of the Interim Procedures and the results of these estimates should be documented in the form of isopleth maps and tables as well as descriptive material that interprets the results.

Finding 9. Detailed performance evaluation objectives were generally not established before writing the protocols.

Recommendation

It is believed that the development and submission of detailed performance evaluation objectives should lead to logical and perhaps more uniform choices of performance measures, averaging times, pairing and weighting in the protocol. Then the rationale for the choices will be explicit to the reviewer, and this should facilitate any negotiation. Thus the use of detailed performance evaluation objectives is encouraged.

Finding 10. A wide variation in the choice of data sets, averaging times, pairing, performance measures, and weighting is evident among the protocols.

Recommendation

While EPA is not necessarily concerned about these wide variations at this time, it is important that the rationale for the choices be documented; the recommendation regarding performance evaluation objectives, above, is one way to establish this rationale.

Finding 11. Similarly, a wide variety in the schemes used for objectively determining the degree to which the models reproduce the measured concentration (scoring) is evident.

Recommendation

Same as Item (10) above.

Finding 12. More recent protocols contain stipulations for adjusting estimates from the chosen model, should that model be shown to underestimate critical concentrations.

Recommendation

The use of model "adjustment factors" to take care of model underestimates was a result of EPA's concerns. While the "adjustment factor" approach is acceptable for the time being, the development of more innovative and more scientifically defensible schemes to address underestimates is encouraged.

Finding 13. The data bases to be used in the performance evaluations consist of networks of 7 to 11 monitors, primarily clustered in the area of expected maximum concentration. Meteorological data from on-site towers are generally to be used in the evaluations.

Recommendation

These limited monitoring networks were acceptable because the areal and temporal extent of the critical source-receptor relationships in the five protocols was very limited. In many cases it may not be possible to establish a priori these critical source-receptor relationships. In such cases more monitors might be required.

The need for representative meteorological data is critical to the performance of the models. To ensure that this need is met, the practice of collecting on-site meteorological data, commensurate with the model's input requirements, is encouraged.

Finding 14. Emissions data are either derived from in-stack instrumentation (two cases) or from surrogate measures such as fuel samples, load levels, etc. (three cases).

Recommendation

The use of surrogate data such as fuel sampling, load levels, etc. leads to considerable uncertainty in emissions especially when coal fired boilers or industrial process emissions are involved. The use of continuous in-stack instrumentation is encouraged.

5.0 REFERENCES

1. Environmental Protection Agency. "Guideline on Air Quality Models," EPA-450/2-78-027, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, April 1978.

2. Environmental Protection Agency. "Interim Procedures for Evaluating Air Quality Models (Revised)," EPA-450/4-84-023, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, September 1984.

3. Fox. D. G. "Judging Air Quality Model Performance," Bull. Am. Meteor. Soc. 62, 599-609, May 1981.

4. Illinois Power and the Illinois Environmental Protection Agency. "Procedures for Model Evaluation and Emission Limit Determination for the Baldwin Power Plant," June 1982.

5. Environmental Protection Agency. "Workbook for Comparison of Air Quality Models," EPA 450/2-78-028a,b, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, May 1978.

6. Environmental Research & Technology, Inc. "Evaluation of MPSDM and CRSTER using the Illinois EPA-approved Protocol and the Subsequent Emission Limitation Study for the Baldwin Power Plant," Documents P-B881-100, P-B881-200, Prepared for Illinois Power Company, July 1983, July 1984

7. Hanna, S., C. Vaudo, A. Curreri, J. Beebe, B. Egan, and J. Mahoney. "Diffusion Model Development and Evaluation and Emission Limitations at the Westvaco Luke Mill," Document PA439, Prepared for the Westvaco Corporation by Environmental Research and Technology Inc., 696 Virginia Road, Concord, MA 01742, March 1982.

8. Bowers, J. F., H. E. Cramer, W. R. Hargraves and A. J. Anderson. "Westvaco Luke, Maryland Monitoring Program: Data Analysis and Dispersion Model Validation," Final Report prepared for U.S. Environmental Protection Agency, Region III by H.E. Cramer Company Inc., University of Utah Research Park, Post Office Box 8049, Salt Lake City, UT 84108, June 1983.

9. Hanna, Steven B., Bruce A. Egan, Cosmos J. Vaudo and Anthony J. Curreri. "An Evaluation of the LUMM and SHORTZ Dispersion Models Using the Westvaco Data Set," Document PA-439, Prepared for the Westvaco Corporation by Environmental Research & Technology, Inc., 696 Virginia Road, Concord, MA 01742, November 1982.

10. Londergan, Richard J. "Protocol for the Comparative Performance Evaluation of the LAPPES and Complex I Dispersion Models Using the Warren Data Set," TRC Environmental Consultants, Inc., 800 Connecticut Blvd., East Hartford, CT 06108, November 1984.

11. Environmental Protection Agency. "Regional Workshop on Air Quality Modeling: A Summary Report," EPA 450/4-82-015, Office of Air Quality Planning and Standards, Research Triangle Park, NC 27711, April 1981.

12. Environmental Research & Technology, Inc. "Protocol for the Evaluation and Comparison of Air Quality Models for Lovett Generating Station," Document P-B636-100, Prepared for Orange & Rockland Utilites, Inc., July 1984.

13. Environmental Protection Agency. Letter to Mr. Frank E. Fischer, Vice President, Engineering, Orange & Rockland Utilites, Inc., EPA, Region II, 26 Federal Plaza, New York, NY 10278, August 30, 1984.

14. Environmental Research & Technology, Inc. "Validation of the Puerto Rico Air Quality Model for the Guayanilla Basin," Document P-9050, Prepared for Environmental Quality Board of Puerto Rico, November 1979.

15. Environmental Protection Agency. "Protocol for the Comparative Performance Evaluation of the PRAQM and Complex I Dispersion Models in the Guayanilla Basin," EPA Region II, 26 Federal Plaza, New York, NY 10278, August 1984.

16. Public Service Company of Indiana. "Plan for Field Study Leading to Model Evaluation and Emission Limit Determination for the Gibson Generating Station," Document P-A892, Environmental Research & Technology, Inc., 696 Virginia Road, Concord, MA 01742, May 1981.

17. Environmental Protection Agency. Letter to Mr. S. A. Ali, Public Service Indiana from Environmental Protection Agency, Region V, 230 South Dearborn Street, Chicago, IL 60604, August 10, 1982.

18. Environmental Protection Agency. Letter to Mr. S. A. Ali, Public Service Indiana from Environmental Protection Agency, Region V, 230 South Dearborn Street, Chicago, IL 60604, June 10, 1982.

19. Burkhart, Richard P. "Protocol for the Comparative Performance Evaluation of the LAPPES and Complex I Dispersion Models Using the Penelec Data Set," Pennsylvania Electric Company, 1001 Broad Street, Johnstown, PA 15907, November 15, 1982.

20. Burkhart, Richard P., Richard J. Londergan, Richard A. Rothstein and Herbert S. Borenstein. "Comparative Performance Evaluation of Two Complex Terrain Dispersion Models," Preprint Paper No. 83-47.4, 76th Annual Meeting of the Air Pollution Control Association, June 19-24, 1983.

APPENDIX A

Protocol and Performance Evaluation Results

for

Baldwin Power Plant

PERFORMANCE EVALUATION PROTOCOL AND FINAL SCORES FOR BALDWIN POWER PLANT

DATA SET	PAIRING		PERFORMANCE MEASURES	AVERAGING TIMES	MAXIMUM POINTS	SCORING SCHEME (CODE)*	WEIGHTING		SCORES	
	SPACE	TIME					INDIV-IDUAL	DATA SET	MPSDM	CRSTER
Second-	Yes	Yes	d	3-hour	15	a	15	55	0.0	0.0
Highest	No	No	d	3-hour	40	a	40		17.7	14.0
25-	Yes	Yes	$\overline{d}$	3-hour	5	b	5	45	0.1	0.2
Highest	No	No	d	3-hour	15	b	15		13.5	9.0
	Yes	Yes	S_d	3-hour	2.5	c	2.5		1.7	1.8
	No	No	S_d	3-hour	5	c	5		4.4	4.0
	Yes	Yes	$RMSE_d$	3-hour	2.5	c	2.5		1.0	1.1
	No	No	$RMSE_d$	3-hour	5	c	5		4.4	3.6
			No. of Cases In Common	3-hour	5	d	5		4.0	4.0
	Yes	No	Cumulative Frequency Distribution	3-hour	5	e	5		4.5	4.0
						TOTAL	100	100	51.3	41.7

*Letters in this column refer to the specific scoring scheme to be used. See subsequent page(s).

SCORING SCHEME

a. Second-highest data set: Single-valued residuals (d), paired and unpaired

A match between observed and predicted concentration is awarded a maximum skill score, while a residual (observed minus predicted concentration) that is more than 1/2 the observed highest, second-highest concentration in magnitude is assigned a score of zero. Regardless of the sign of the residual, the points awarded vary linearly between 0 and 100% of the maximum possible as the model error varies in magnitude between 1/2 the observed highest, second-highest 3-hour average and zero.

b. 25-highest data set: Bias ($\bar{d}$), paired and unpaired

A scoring scheme for the bias that is the same as that used for the second-high values is used. A zero skill level is assigned to a bias equal to 1/2 of the average observed value for the highest-25 3-hour SO_2 concentrations. The total number of points awarded to a model vary between 0 and the maximum value as the magnitude of the average residual varies between 1/2 the average observed 3-hour concentration and zero.

c. 25-highest data set: Noise and gross variability (S_d, $RMSE_d$), paired and unpaired

The scoring scheme for the noise and gross variability tests involves the ratio of the model precision measure to the average value about which it is being computed. For the noise test, this ratio is the standard deviation divided by the average modeled value. For the gross variability test, the ratio is the root-mean-square-error divided by the coefficient of variation value (standard deviation divided by the mean) often used in statistical testing. A score of 0 points is suggested for a ratio of 1.0, linearly increasing to the maximum score as the ratio goes to zero. That is, the score will be:

SCORE = (1.0 - computed ratio) x the maximum possible points

d. 25-highest data set: Meteorological cases in commmon, unpaired

For the meteorological conditions comparison, 4 general weather categories is used:

1. Unstable (Classes A-C), with the 10-meter wind speed less than 5 m/sec;

2. Neutral (Class D), with the 10-meter wind speed less than 5 m/sec;

3. Stable (Classes E-G), with the 10-meter wind speed less than 5 m/sec;

4. Any case with the 10-meter wind speed greater than 5 m/sec.

The number of cases for each weather category is totaled for the top 25 observed and modeled 3-hour cases. The number of unpaired cases "in common" between observed and predicted 3-hour events is totaled to determine the score for this test for each model:

$$\text{Score} = \left[\frac{\text{No. of Cases in Common}}{25}\right] \left[\text{Maximum Points}\right]$$

e. 25-highest data set: Cumulative frequency distribution, paired in space

For each individual monitor, the Kolomogorov-Smirnoff (K-S) test is be used to determine whether the cumulative frequency distributions between the top 25 observed and predicted 3-hour values are significantly different (at the 5% significance level). Points are awarded for each monitor for which there is not a significant difference in a cumulative frequency distribution:

$$\text{Score} = \left[\frac{\text{No. of monitors where frequency distributions are not significantly different}}{25}\right] \left[\text{Maximum Points}\right]$$

APPENDIX B

Protocol and Performance Evaluation Results

for

Westvaco Luke Mill

PERFORMANCE EVALUATION PROTOCOL AND FINAL SCORES FOR WESTVACO LUKE MILL

DATA SET	PAIRING		PERFORMANCE MEASURES	AVERAGING TIMES	MAXIMUM POINTS	SCORING SCHEME (CODE)*	WEIGHTING		SCORES	
	SPACE	TIME					INDIV-IDUAL	DATA SET	LUMM	SHORTZ
Maximum	No	No	|d|	3-hour	20	a	3.3	21.2	12	0
				24-hour	20	a	3.3		19	0
	Yes	No	|d| for 8 monitors	3-hour	16(1)	a	2.7		10	4
				24-hour	16(1)	a	2.7		13	3
			|d| for monitor #10	3-hour	8	a	1.3		0	8
				24-hour	8	a	1.3		1	8
	No	Yes	|d|	Annual	20	a	3.3		13	1
	Yes	Yes	|d|	Annual	20	a	3.3		9	8
Second-Highest	No	No		3-hour	30	a	5.0	22.0	27	0
				24-hour	30	a	5.0		26	0
	Yes	No	|d| for 8 monitors	3-hour	24(2)	a	4.0		17	5
				24-hour	24(2)	a	4.0		4	10
			|d| for monitor #10	3-hour	12	a	2.0		16	6
				24-hour	12	a	2.0		3	11
25-Highest	No	No	|d|	1-hour	25	b	4.2	56.7	17	0
				3-hour	25	b	4.2		23	0
				24-hour	25	b	4.2		23	0
			$s_p^2/s_o^2, s_o^2/s_p^2$	1-hour	5	c	0.8		0	0
				3-hour	5	c	0.8		1	0
				24-hour	5	c	0.8		4	0
	Yes	No	|d| for 8 monitors	1-hour	40(3)	b	6.6		30	5
				3-hour	40(3)	b	6.6		30	8
				24-hour	40(3)	b	6.6		26	9
			$s_p^2/s_o^2, s_o^2/s_p^2$ for 8 monitors	1-hour	16(4)	c	2.7		6	2
				3-hour	16(4)	c	2.7		5	4
				24-hour	16(4)	c	2.7		7	4
			|d| for monitor #10	1-hour	20	b	3.3		1	19
				3-hour	20	b	3.3		9	16
				24-hour	20	b	3.3		4	19
			$s_p^2/s_o^2, s_o^2/s_p^2$ for monitor #10	1-hour	8	c	1.3		2	2
				3-hour	8	c	1.3		2	8
				24-hour	8	c	1.3		3	8
				TOTAL	602		99.9(5)	99.9(5)	363	168

*Letters in this column refer to the specific scoring scheme used. See subsequent page(s).

Footnotes:

(1) 2 points per monitor
(2) 3 points per monitor
(3) 5 points per monitor
(4) 2 points per monitor
(5) Do not add to 100% because of rounding

SCORING SCHEME

a. Maximum and second-highest data sets: Singled-valued residual ($|d|$), various pairings

$$\text{Score} = [|d|_{min}/|d|_i]\ [\min(C_{p,i}/C_o, C_o/C_{p,i})]\ [\text{max points}]$$

Where $i = 1,2$ = Model 1 or Model 2

b. 25-highest data set: Bias ($|d|$), unpaired, paired in space

$$\text{Score} = [|d|_{min}/|d|_i]\ [\min(C_{p,i}/C_o, C_o/C_{p,i})]$$

where $i = 1,2$ = Model 1 or Model 2

c. 25-highest data set: Variance ($S_p^2/S_o^2, S_o^2/S_p^2$), unpaired, paired in space

$$\text{Score} = [\min(S_{p,i}^2/S_o^2, S_o^2/S_{p,i}^2)]\ [\text{max points}]$$

where $i = 1,2$ = Model 1 or Model 2

APPENDIX C

Protocol

for

Warren Power Plant

PERFORMANCE EVALUATION PROTOCOL FOR WARREN POWER PLANT

DATA SET	PAIRING		PERFORMANCE MEASURES	AVERAGING TIMES	MAXIMUM POINTS	SCORING SCHEME (CODE)*	WEIGHTING	
	SPACE	TIME					INDIV-IDUAL	DATA SET
Maximum	Yes	No	$\overline{d}$	1-hour	2.0	a	1.4	14.9
				3-hour	2.7	a	1.9	
				24-hour	3.6	a	2.6	
			$\lvert\overline{d}\rvert/\overline{C}_o$	1-hour	2.0	b	1.4	
				3-hour	2.7	b	1.9	
				24-hour	3.6	b	2.6	
			R	1-hour	1.0	c	0.7	
				3-hour	1.6	c	1.1	
				24-hour	1.8	c	1.3	
Second-Highest	No	No	C_p/C_o	3-hour	7.0	d	5.0	28.4
				24-hour	9.0	d	6.4	
	Yes	No	C_p/C_o (1)	3-hour	12.0	e	8.5	
				24-hour	12.0	e	8.5	
25-Highest	No	No	$\overline{C}_p/\overline{C}_o$	1-hour	8.0	f	5.7	42.6
				3-hour	10.0	f	7.1	
				24-hour	13.0	f	9.2	
			s_p^2/s_o^2	1-hour	4.0	g	2.8	
				3-hour	6.0	g	4.3	
				24-hour	7.0	g	5.0	
			$\overline{C}_p/\overline{C}_o$ (2)	1-hour	8.0	h	5.7	
			s_p^2/s_o^2 (2)	1-hour	4.0	i	2.8	
All Data	No	No	$\overline{C}_p/\overline{C}_o$	Annual	8.0	j	5.7	14.1
	Yes	No	$\overline{d}$	Annual	4.0	k	2.8	
			$\lvert\overline{d}\rvert/\overline{C}_o$	Annual	4.0	l	2.8	
			R	Annual	4.0	m	2.8	
			TOTAL		141		100.0	100.0

*Letters refer to the specific scoring scheme to be used. See subsequent page(s).

Footnotes:

(1) For stations with the 3 highest observed and 3 highest estimated values--see scoring scheme below.

(2) Stratified by stability--see scoring scheme below.

SCORING SCHEME

a. Maximum data set: Average difference ($\bar{d}$), paired in space

Confidence intervals for 50 percent, 80 percent, 95 percent confidence levels from t-test.

	Point Score		
	(1-Hour)	(3-Hour)	(24-Hour)
50 percent confidence interval (C.I.) contains zero (observed=predicted)	2.0	2.7	3.6
80 percent C.I. contains zero (but 50 percent does not)	1.33	1.8	2.4
95 percent C.I. contains zero (but 80 percent does not)	0.67	0.9	1.2
95 percent C.I. does not contain zero	0	0	0

b. Maximum data set: Average absolute difference (AAD), paired in space

Compute ratio of AAD to average observed value.

	Point Score		
	(1-Hour)	(3-Hour)	(24-Hour)
ratio $\leq$ 0.2	2	2.7	3.6
0.2 < ratio $\leq$ 0.4	1.33	1.8	2.4
0.4 < ratio $\leq$ 0.8	0.67	0.9	1.2
0.8 < ratio	0	0	0

c. Maximum data set: Pearson's correlation coefficient (R), paired in space

	Point Score		
	(1-Hour)	(3-Hour)	(24-Hour)
R > 0.8	1	1.6	1.8
0.8 $\geq$ R > 0.6	0.5	0.8	0.9
0.6 $\geq$ R	0	0	0

d. Second-highest data set: Highest second-highest value, unpaired

C_p/C_o = the ratio of the predicted to the observed highest second-high value.

	Point Score	
	3-hour	24-hour
$0.5 > C_p/C_o$	0	0
$0.67 > C_p/C_o \geq 0.5$	2.6	3.4
$0.83 > C_p/C_o \geq 0.67$	4.4	5.6
$1.2 > C_p/C_o \geq 0.83$	7	9
$1.5 > C_p/C_o \geq 1.2$	4.4	5.6
$2.0 > C_p/C_o \geq 1.5$	2.6	3.4
$C_p/C_o \geq 2.0$	0	0

e. Second-highest data set: Second-highest observed and predicted value (by stations with the highest, second-highest, and third-highest values (12 points possible), paired in space

C_p/C_o = ratio of predicted to observed second-highest value at the same station

	Point Score		
	Station w/highest value	Second-highest station	Third-highest station
$0.5 > C_p/C_o$	0	0	0
$0.67 > C_p/C_o \geq 0.5$	1	1	0
$0.83 > C_p/C_o \geq 0.67$	2	1	1
$1.2 > C_p/C_o \geq 0.83$	3	2	1
$1.5 > C_p/C_o \geq 1.2$	2	1	1
$2.0 > C_p/C_o \geq 1.5$	1	1	0
$C_p/C_o \geq 2.0$	0	0	0

f. 25-highest data set: bias ($\overline{C}_p/\overline{C}_o$), unpaired

$\overline{C}_p/\overline{C}_o$ = ratio of predicted to observed average value

	Point Score		
	(1-hour)	(3-hour)	(24-hour)
$0.67 > \overline{C}_p/\overline{C}_o$	0	0	0
$0.83 > \overline{C}_p/\overline{C}_o \geq 0.67$	2.5	3	4
$0.91 > \overline{C}_p/\overline{C}_o \geq 0.83$	5	6	8
$1.1 > \overline{C}_p/\overline{C}_o \geq 0.91$	8	10	13
$1.2 > \overline{C}_p/\overline{C}_o \geq 1.1$	5	6	8
$1.5 > \overline{C}_p/\overline{C}_o \geq 1.2$	2.5	3	4
$\overline{C}_p/\overline{C}_o \geq 1.5$	0	0	0

g. 25-highest data set: variance ratio (S_p^2/S_o^2), unpaired

S_p^2/S_o^2 = ratio of predicted to observed variance

	Point Score		
	(1-hour)	(3-hour)	(24-hour)
$0.25 > S_p^2/S_o^2$	0	0	0
$0.50 > S_p^2/S_o^2 \geq 0.25$	1.33	2	2.4
$0.75 > S_p^2/S_o^2 \geq 0.50$	2.67	4	4.8
$1.33 > S_p^2/S_o^2 \geq 0.75$	4	6	7
$2.0 > S_p^2/S_o^2 \geq 1.33$	2.67	4	4.8
$4.0 > S_p^2/S_o^2 \geq 2.0$	1.33	2	2.4
$S_p^2/S_o^2 \geq 4.0$	0	0	0

h. 25-highest data set: Bias ($\overline{C}_p/\overline{C}_o$), by stability category, unpaired

For the stability category with the highest observed concentrations, compare the 25-highest observed and 25-highest predicted values (unpaired in time or location). Repeat for the stability category with the highest predicted concentrations. (1-hour average only).

Ratio of predicted to observed average value

	Point Score
$0.67 > \overline{C}_p/\overline{C}_o$	0
$0.83 > \overline{C}_p/\overline{C}_o \geq 0.67$	1.25
$0.91 > \overline{C}_p/\overline{C}_o \geq 0.83$	2.5
$1.1 > \overline{C}_p/\overline{C}_o \geq 0.91$	4
$1.2 > \overline{C}_p/\overline{C}_o \geq 1.1$	2.5
$1.5 > \overline{C}_p/\overline{C}_o \geq 1.2$	1.25
$\overline{C}_p/\overline{C}_o \geq 1.5$	0

i. 25-highest data set: Variance ratio (s_p^2/s_o^2), by stability category, unpaired

For the stability category with the highest observed concentrations, compare the 25 highest observed and 25 highest predicted values (unpaired in time or location). Repeat for the stability category with the highest predicted concentrations. (1-hour average only).

s_p^2/s_o^2 = Ratio of predicted to observed variance

	Point Score
$0.25 > s_p^2/s_o^2$	0
$0.50 > s_p^2/s_o^2 \geq 0.25$	0.67
$0.75 > s_p^2/s_o^2 \geq 0.50$	1.33
$1.33 > s_p^2/s_o^2 \geq 0.75$	2
$2.0 > s_p^2/s_o^2 \geq 1.33$	1.33
$4.0 > s_p^2/s_o^2 \geq 2.0$	0.67
$s_p^2/s_o^2 \geq 4$	0

j. All data set: Bias ($\overline{C}_p/\overline{C}_o$), unpaired

Ratio of predicted to observed highest value

	Point Score
$0.75 > \overline{C}_p/\overline{C}_o$	0
$0.83 > \overline{C}_p/\overline{C}_o \geq 0.75$	2
$0.91 > \overline{C}_p/\overline{C}_o \geq 0.83$	4
$0.95 > \overline{C}_p/\overline{C}_o \geq 0.91$	6
$1.05 > \overline{C}_p/\overline{C}_o \geq 0.95$	8
$1.1 > \overline{C}_p/\overline{C}_o \geq 1.05$	6
$1.2 > \overline{C}_p/\overline{C}_o \geq 1.1$	4
$1.33 > \overline{C}_p/\overline{C}_o \geq 1.2$	2
$\overline{C}_p/\overline{C}_o \geq 1.33$	0

k. All data set: Average residual ($\overline{d}$), paired in space

Use confidence intervals as in a.

	Point Score
50 percent confidence interval contains zero	4
80 percent C. I. contains zero (but 50% does not)	2
95 percent C. I. contains zero (but 80% does not)	1
95 percent C. I. does not contain zero	0

l. All data set: Ratio of average absolute difference to average observed value, paired in space

	Point Score
$0.1 < \lvert\overline{d}\rvert/\overline{C}_o$	4
$0.2 < \lvert\overline{d}\rvert/C_o \leq 0.1$	2
$0.3 < \lvert\overline{d}\rvert/C_o \leq 0.2$	1
$\lvert\overline{d}\rvert/C_o \leq 0.3$	0

m. All data set: Pearson's correlation coefficient (R), paired in space.

	Point Score
$0.9 \geq R$	4
$0.8 \geq R > 0.9$	3
$0.7 \geq R > 0.8$	2
$0.6 \geq R > 0.7$	1
$R > 0.6$	0

APPENDIX D

Protocol

for

Lovett Power Plant

PERFORMANCE EVALUATION PROTOCOL FOR LOVETT POWER PLANT

DATA SET	PAIRING		PERFORMANCE MEASURES	AVERAGING TIMES	MAXIMUM POINTS	SCORING SCHEME (CODE)*	WEIGHTING	
	SPACE	TIME					INDIV-IDUAL	DATA SET
Second Highest	No	No	C_p/C_o, C_o/C_p	3-hour	5.0	a	5.0	20.0
				24-hour	5.0	a	5.0	
	Yes	No	C_p/C_o, C_o/C_p	3-hour	5.0	b	5.0	
				24-hour	5.0	b	5.0	
25-Highest	No	No		1-hour	4.0	c	4.0	58.0
			$\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$	3-hour	4.0	c	4.0	
				24-hour	4.0	c	4.0	
	Yes	No		1-hour	4.0	d	4.0	
			$\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$	3-hour	4.0	d	4.0	
				24-hour	4.0	d	4.0	
	No	No	S_p^2/S_o^2,	1-hour	4.0	e	4.0	
				3-hour	4.0	e	4.0	
			S_o^2/S_p^2	24-hour	4.0	e	4.0	
	Yes	No	S_p^2/S_o^2,	1-hour	4.0	f	4.0	
				3-hour	4.0	f	4.0	
			S_o^2/S_p^2	24-hour	4.0	f	4.0	
	No	No	No. of cases in common	1-hour	10.0	g	10.0	
All Data	No	Yes	$\overline{C}_p/\overline{C}_o$, $\overline{C}_p/\overline{C}_p$	Annual	1.0	h	1.0	22.0
	Yes	Yes	$\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$	Annual	1.0	i	1.0	
	No	Yes	$R^{(1)}$	1-hour	1.0	j	1.0	
				3-hour	1.0	j	1.0	
			$\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p{}^{(1)}$	1-hour	1.0	k	1.0	
				3-hour	1.0	k	1.0	
			S_p^2/S_o^2	1-hour	1.0	l	1.0	
			$S_o^2/S_p^{2\,(1)}$	3-hour	1.0	l	1.0	
			$d^{2(1)}$	1-hour	2.0	m	2.0	
				3-hour	2.0	m	2.0	
			$R^{(2)}$	1-hour	1.0	n	1.0	
				3-hour	1.0	n	1.0	
			$\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p{}^{(2)}$	1-hour	1.0	o	1.0	
				3-hour	1.0	o	1.0	
			S_p^2/S_o^2, $S_o^2/S_p^{2(2)}$	1-Hour	1.0	p	1.0	
				3-hour	1.0	p	1.0	
			$d^{2(2)}$	1-hour	2.0	q	2.0	
				3-hour	2.0	q	2.0	
				TOTAL	100.0		100.0	100.0

*Letters refer to specific scoring scheme to be used. See subsequent page(s).

Footnotes: (1) Stable conditions
(2) Nonstable conditions

SCORING SCHEME

a. Second-highest data set: Ratios of concentrations (C_p/C_o, C_o/C_p), unpaired

Score = [min(C_p/C_o, C_o/C_p)] [max points]

b. Second-highest data set: Ratios of concentrations (C_p/C_o, C_o/C_p), paired in space

Score = [min (C_p/C_o, C_o/C_p)] [max points]

c. 25-highest data set: Bias ($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$), unpaired

Score = [min ($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$)] [max points]

d. 25-highest data set: Bias ($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$), paired in space

Score = [min ($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$)] [max points]

e. 25-highest data set: Variance ratios (S_p^2/S_o^2, S_o^2/S_p^2), unpaired

Score = [min (S_p^2/S_o^2, S_o^2/S_p^2)] [max points]

f. 25-highest data set: Variance ratios (S_p^2/S_o^2, S_o^2/S_p^2), paired in space

Score = [min (S_p^2/S_o^2, S_o^2/S_p^2)] [max points]

g. 25-highest data set: Number of meteorological cases in common

$$\text{Score} = \left[\frac{\text{No. of cases in common}}{25}\right] \text{[max points]}$$

h. All data set: Bias ($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$), paired in space

Score = [min($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$)] [max points]

i. All data set: Bias ($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$), paired in space and time

Score = [min($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$)] [max points]

j. All data set: Pearson's correlation coefficient (R), stable conditions only, paired in time

Score = [R^2] [max points]

k. All data set: Bias ($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$), stable conditions only, paired in time

Score = [min($\overline{C}_p/\overline{C}_o$, $\overline{C}_o/\overline{C}_p$)] [max points]

l. All data set: Variance ratios (S_p^2/S_o^2, S_o^2/S_p^2), stable conditions only, paired in time

Score = [min(S_p^2/S_o^2, S_o^2/S_p^2)] [max points]

m. All data set: Gross variability (d^2), stable conditions only, paired in time

Score = [$(\sum d^2)_{min}/(\sum d^2)$] [max points]

where $(\sum d^2)_{min}$ = $\sum d^2$ for best performing model

n. All data set: Pearson's correlation coefficient (R), nonstable conditions, paired in time

Score = [R^2] [max points]

o. All data set: Bias ($\overline{C}_p/\overline{C}_o$,$\overline{C}_o/\overline{C}_p$), nonstable conditions, paired in time

Score = [min($\overline{C}_p/\overline{C}_o$,$\overline{C}_o/\overline{C}_p$)] [max points]

p. All data set: Variance ratios (S_p^2/S_o^2, S_o^2/S_p^2), nonstable conditions, paired in time

Score = [min(S_p^2/S_o^2, S_o^2/S_p^2)] [max points]

q. All data set: Gross variability (d^2), nonstable conditions, paired in time

Score = [$(\sum d^2)_{min}/(\sum d^2)$] [max points]

where $(\sum d^2)_{min}$ = $\sum d^2$ for best performing model

APPENDIX E

Protocol

for

Guayanilla Basin

PERFORMANCE EVALUATION PROTOCOL FOR GUAYANILLA BASIN

DATA SET	PAIRING		PERFORMANCE MEASURES	AVERAGING TIMES	MAXIMUM POINTS	SCORING SCHEME (CODE)*	WEIGHTING	
	SPACE	TIME					INDIV-IDUAL	DATA SET
Maximum	No	No	C_p/C_o	3-hour	4	a	1.6	11.8
				24-hour	5	a	2.0	
	Yes	No	C_p/C_o	3-hour	9	b	3.5	
				24-hour	12	b	4.7	
Second-Highest	No	No	C_p/C_o	3-hour	5	c	2.0	14.9
				24-hour	6	c	2.3	
	Yes	No	C_p/C_o	3-hour	12	d	4.7	
				24-hour	15	d	5.9	
25-Highest	No	No	$\overline{C}_p/\overline{C}_o$	1-hour	6	e	2.3	50.2
				3-hour	8	e	3.1	
				24-hour	12	e	4.7	
			S_p^2/S_o^2	1-hour	3	f	1.2	
				3-hour	4	f	1.6	
				24-hour	6	f	2.3	
	Yes	No	$\overline{C}_p/\overline{C}_o$	1-hour	12	g	4.7	
				3-hour	18	g	7.0	
				24-hour	30	g	11.7	
			S_p^2/S_o^2	1-hour	6	h	2.3	
				3-hour	9	h	3.5	
				24-hour	15	h	5.8	
Upper 5% of Observed & Predicted	Yes	No	No. of Cases in Common	1-hour	60	i	23.4	23.4
				TOTAL	257		100.3 (1)	100.3 (1)

*Letters refer to specific scoring scheme to be used. See subsequent page(s).

Footnote:
(1) Do not add to 100% because of rounding.

SCORING SCHEME

a. Maximum data set: Concentration ratio (C_p/C_o), unpaired

	3-hr	24-hr
$0.67 > C_p/C_o$	0.0	0.0
$0.80 > C_p/C_o \geq 0.67$	0.5	1.0
$0.91 > C_p/C_o \geq 0.80$	2.0	2.5
$1.20 > C_p/C_o \geq 0.91$	4.0	5.0
$1.50 > C_p/C_o \geq 1.20$	2.5	3.5
$2.50 > C_p/C_o \geq 1.50$	1.5	2.0
$C_p/C_o \geq 2.50$	0.0	0.0

b. Maximum data set: Concentration ratio (C_p/C_o), paired in space

A weighting factor is to be applied to the scores for the tests at each monitor. The weighting factor will be based on the relative rank of the observed data for each averaging period to be examined. The following weights will be assigned and should be applied to the table below.

Phase I		Phase II	
Rank	Weight	Rank	Weight
1	1.0	1,2	0.50
2	0.8	3,4	0.40
3	0.7	5,6	0.35
4	0.5	7,8	0.25

	3-hr	24-hr
$0.67 > C_p/C_o$	0.0	0.0
$0.80 > C_p/C_o \geq 0.67$	0.0	0.5
$0.91 > C_p/C_o \geq 0.80$	1.0	2.0
$1.20 > C_p/C_o \geq 0.91$	3.0	4.0
$1.50 > C_p/C_o \geq 1.20$	2.0	2.5
$2.50 > C_p/C_o \geq 1.50$	1.0	1.5
$C_p/C_o \geq 2.50$	0.0	0.0

c. Second-highest data set: Concentration ratio (C_p/C_o), unpaired

	3-hr	24-hr
$0.67 > C_p/C_o$	0.0	0.0
$0.80 > C_p/C_o \geq 0.67$	1.0	1.5
$0.91 > C_p/C_o \geq 0.80$	2.5	3.0
$1.20 > C_p/C_o \geq 0.91$	5.0	6.0
$1.50 > C_p/C_o \geq 1.20$	3.5	4.0
$2.50 > C_p/C_o \geq 1.50$	2.0	2.5
$C_p/C_o \geq 2.50$	0.0	0.0

d. Second-highest data set: Concentration ratio (C_p/C_o), paired in space

A weighting factor is to be applied to the scores for the tests at each monitor. The weighting factor will be based on the relative rank of the observed data for each averaging period to be examined. The following weights will be assigned and should be applied to the table below.

Phase I		Phase II	
Rank	Weight	Rank	Weight
1	1.0	1,2	0.50
2	0.8	3,4	0.40
3	0.7	5,6	0.35
4	0.5	7,8	0.25

	3-hr	24-hr
$0.67 > C_p/C_o$	0.0	0.0
$0.80 > C_p/C_o \geq 0.67$	0.5	1.0
$0.91 > C_p/C_o \geq 0.80$	2.0	2.5
$1.20 > C_p/C_o \geq 0.91$	4.0	5.0
$1.50 > C_p/C_o \geq 1.20$	2.5	3.5
$2.50 > C_p/C_o \geq 1.50$	1.5	2.0
$C_p/C_o \geq 2.50$	0.0	0.0

e. 25 highest data set: Bias ($\bar{C}_p/\bar{C}_o$), unpaired

	1-hr	3-hr	24-hr
$0.67 > \bar{C}_p/\bar{C}_o$	0.0	0.0	0.0
$0.80 > \bar{C}_p/\bar{C}_o \geq 0.67$	1.5	2.0	3.0
$0.91 > \bar{C}_p/\bar{C}_o \geq 0.80$	3.0	4.0	6.0
$1.20 > \bar{C}_p/\bar{C}_o \geq 0.91$	6.0	8.0	12.0
$1.50 > \bar{C}_p/\bar{C}_o \geq 1.20$	4.0	5.5	8.0
$2.50 > \bar{C}_p/\bar{C}_o \geq 1.50$	2.5	3.0	4.0
$\bar{C}_p/\bar{C}_o \geq 2.50$	0.0	0.0	0.0

f. 25 highest data set: Variance ratio (S_p^2/S_o^2), unpaired

	1-hr	3-hr	24-hr
$S_p^2/S_o^2 \leq 0.25$	0.0	0.0	0.0
$0.25 < S_p^2/S_o^2 \leq 0.50$	1.0	1.5	2.0
$0.50 < S_p^2/S_o^2 \leq 0.75$	2.0	3.0	4.0
$0.75 < S_p^2/S_o^2 \leq 1.33$	3.0	4.0	6.0
$1.33 < S_p^2/S_o^2 \leq 2.00$	2.0	3.0	4.0
$2.00 < S_p^2/S_o^2 \leq 4.00$	1.0	1.5	2.0
$4.00 < S_p^2/S_o^2$	0.0	0.0	0.0

g. 25 highest data set: Bias ($\overline{C}_p/\overline{C}_o$), paired in space

A weighting factor is to be applied to the scores for the tests at each monitor. The weighting factor will be based on the relative rank of the observed data for each averaging period to be examined. The following weights will be assigned and should be applied to the table below.

Phase I		Phase II	
Rank	Weight	Rank	Weight
1	1.0	1,2	0.50
2	0.8	3,4	0.40
3	0.7	5,6	0.35
4	0.5	7,8	0.25

	1-hr	3-hr	24-hr
$0.67 > \overline{C}_p/\overline{C}_o$	0.0	0.0	0.0
$0.80 > \overline{C}_p/\overline{C}_o \geq 0.67$	0.5	1.5	2.5
$0.91 > \overline{C}_p/\overline{C}_o \geq 0.80$	2.0	3.0	5.0
$1.20 > \overline{C}_p/\overline{C}_o \geq 0.91$	4.0	6.0	10.0
$1.50 > \overline{C}_p/\overline{C}_o \geq 1.20$	2.5	4.0	6.5
$2.50 > \overline{C}_p/\overline{C}_o \geq 1.50$	1.5	2.5	3.5
$\overline{C}_p/\overline{C}_o \geq 2.50$	0.0	0.0	0.0

h. 25 highest data set: Variance ratio (S_p^2/S_o^2), paired in space

A weighting factor is to be applied to the scores for the tests at each monitor. The weighting factor will be based on the relative rank of the observed data for each averaging period to be examined. The following weights will be assigned and should be applied to the table below.

Phase I		Phase II	
Rank	Weight	Rank	Weight
1	1.0	1,2	0.50
2	0.8	3,4	0.40
3	0.7	5,6	0.35
4	0.5	7,8	0.25

	1-hr	3-hr	24-hr
$s_p^2/s_o^2 \leq 0.25$	0.0	0.0	0.0
$0.25 < s_p^2/s_o^2 \leq 0.50$	0.50	1.0	2.0
$0.50 < s_p^2/s_o^2 \leq 0.75$	1.0	2.0	3.5
$0.75 < s_p^2/s_o^2 \leq 1.33$	2.0	3.0	5.0
$1.33 < s_p^2/s_o^2 \leq 2.00$	1.0	2.0	3.5
$2.00 < s_p^2/s_o^2 \leq 4.00$	0.5	1.0	2.0
$4.00 < s_p^2/s_o^2$	0.0	0.0	0.0

i. Upper 5% of frequency distribution data set: Number of cases in common

At each monitor, unpaired in time, stratify the upper 5% of the 1-hour predicted and observed concentrations according to the following categories:

I. Unstable (Classes A, B, C)
II. Neutral (Class D)
III. Stable (Classes E, F)

The number of unpaired cases "in common" between observed and predicted 1-hour events will be used to determine a skill factor for each category, defined as:

$$R_{sf} = \frac{2 \times (\text{Number in Common})}{(\text{Number Predicted} + \text{Number Observed})}$$

The total number of points for each Phase I monitor is 15 points and total points for each Phase II monitor is 7.5 points, appropriated as follows:

from the highest 25 observed concentrations,

	Phase I	Phase II
Most predominant category:	8 pts.	4 pts.
Next predominant category:	4 pts.	2 pts.
Least predominant category:	3 pts.	1.5 pts.

The total score is given by

$$\text{Score} = \sum_{\text{categories}} (R_{sf})(\text{max points})$$

TECHNICAL REPORT DATA
(Please read Instructions on the reverse before completing)

1. REPORT NO.	2.	3. RECIPIENT'S ACCESSION NO.
EPA 450/4-85-006		

4. TITLE AND SUBTITLE	5. REPORT DATE
Interim Procedures for Evaluating Air Quality Models: Experience with Implementation	July 1985
	6. PERFORMING ORGANIZATION CODE
7. AUTHOR(S)	8. PERFORMING ORGANIZATION REPORT NO.
9. PERFORMING ORGANIZATION NAME AND ADDRESS Monitoring and Data Analysis Division Office of Air Quality Planning and Standards U. S. Environmental Protection Agency Research Triangle Park, N.C. 27711	10. PROGRAM ELEMENT NO.
	11. CONTRACT/GRANT NO.
12. SPONSORING AGENCY NAME AND ADDRESS	13. TYPE OF REPORT AND PERIOD COVERED
	14. SPONSORING AGENCY CODE
15. SUPPLEMENTARY NOTES	

16. ABSTRACT This report summarizes and intercompares the details of five major regulatory cases for which guidance provided in the "Interim Procedures for Evaluating Air Quality Models" was implemented in evaluating candidate models. In two of the cases the evaluations have been completed and the appropriate model has been determined. In three cases the data base collection and/or the final analysis has not yet been completed. The purpose of the report is to provide potential users of the Interim Procedures with a description and analysis of several applications that have taken place. With this information in mind the user should be able to: (1) more effectively implement the procedures since some of the pitfalls experienced by the initial pioneers can now be avoided; and (2) design innovative technical criteria and statistical techniques that will advance the state of the science of model evaluation.

The analyses show that the basic principles or framework underlying the Interim Procedures is sound and workable in application. The concept of using the results from a prenegotiated protocol for the performance evaluation has been shown to be an appropriate and workable primary basis for objectively deciding on the best model. Similarly, "up-front" negotiation on what constitutes an acceptable data base network has been established as an acceptable way of promoting objectivity in the evaluation. Preliminary concentration estimates and the need for accurate continuous on-site measurements of the requisite model input data are also important.

17. KEY WORDS AND DOCUMENT ANALYSIS

a. DESCRIPTORS	b. IDENTIFIERS/OPEN ENDED TERMS	c. COSATI Field/Group
Air Pollution Meteorology Mathematical Models Performance Evaluation Statistics	Performance Measures Technical Evaluation	4B 12A

18. DISTRIBUTION STATEMENT	19. SECURITY CLASS (This Report)	21. NO. OF PAGES
Unlimited	Unclassified	
	20. SECURITY CLASS (This page)	22. PRICE
	Unclassified	

EPA Form 2220-1 (Rev. 4-77) PREVIOUS EDITION IS OBSOLETE